รายงานวิจัยฉบับสมบูรณ์

โครงการ
"การรวมข้อมูลระหว่างการวิจัยเชิงทดลองสุ่มแบบกลุ่มกับการวิจัยเชิงทดลองสุ่มแบบธรรมดาในการวิเคราะห์เมตต้า"

โดย

รองศาสตราจารย์ ดร.มาลินี เหล่าไพบูลย์

30 กันยายน 2546

สัญญาเลขที่ BRG /15/2543

รายงานวิจัยฉบับสมบูรณ์

โครงการ

"การรวมข้อมูลระหว่างการวิจัยเชิงทดลองสุ่มแบบกลุ่มกับการวิจัยเชิงทดลองสุ่มแบบธรรมดาในการวิเคราะห์เมตต้า"

โดย

รองศาสตราจารย์ ดร.มาลินี เหล่าไพบูลย์

ภาควิชาชีวสถิติและประชากรศาสตร์

คณะสาธารณสุขศาสตร์

มหาวิทยาลัยขอนแก่น

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย

Abstract

Rationale: Most of statistical methods used in meta-analysis assume individual subjects as units of randomization. Meta-analyses involving cluster randomized trials may lead to additional sources of heterogeneity beyond those elevated by meta-analyses involving only individually randomized trials. The appropriate statistical analysis to these meta-analyses must take into account potential heterogeneity in the cluster randomized trials. A substantial amount of literature covering statistical methodologies used in meta analyses can now be found. Most of them, however, assume individual subjects as units of randomization. Therefore, there may remain some questions that need to be investigated in the area of meta analyses related to the inclusion of cluster randomized trials.

The general linear mixed model (GLM) has been proposed to explain heterogeneity in meta-analysis where the treatment effect is measured in binary outcome. Log-relative measure is used as a response variable. The parameter estimation is based on assumption of normal distribution of random effects. The generalized linear mixed model (GLMM) under unspecified distribution of random effects may be an alternative choice. The two approaches allow the inclusion of some covariates of trial level and subject level. Therefore it is interesting to explore potential of the two approaches in meta-analysis involving cluster randomized trials in binary outcome.

Objective: Two potential non-Bayesian approaches of GLM and GLMM are explored to identify and explain heterogeneity in meta-analyses involving cluster randomized trials comparing two treatment groups measured in binary outcome.

Methods: The two approaches of GLM and GLMM are studied and evaluated their potential in term of methodological aspects, results provided, strengths and limitations of these approaches and exemplified in three published meta-analyses involving cluster randomized trials. The first meta-analysis includes eight community-based trials. They were performed in developing countries to examine the relationship of vitamin A supplementation and mortality in children aged 6 to 72 months. None of the trials assigned individual children to treatment groups. The second meta-analysis comprises fewer trials of 8, which is performed to evaluate the effect of mammographic screening on reduction of breast cancer mortality. The third meta-analysis is done to assess the effectiveness of multiple risk factor interventions to reduce cardiovascular risk factors from coronary heart disease. Analysis is performed in the 14 trials included that provided smoking prevalence outcome. For each meta-analysis, observed log-relative risks for individual trials are fitted to the GLM as a continuous response. The trials included are classified to two categories according to randomization units, clusters and individually, and called randomization design variable. This variable is treated as a covariate of the model. The model parameters are estimated with the restricted maximum likelihood (REML) under the normality assumption of random effects via MLwiN software. For the GLMM, observed frequencies of the outcome for each treatment group are used rather than the observed log-relative risks for individual trials. A canonical link function of the observed mean proportions is associated with linear predictors model of which treatment and randomization design are treated as covariates. Here, the treatment effect can be treated as random treatment effects. The maximum likelihood estimates of the model parameters are obtained non-parametrically under a discrete mixture distribution of random effects for K components, which is implemented by the EM-algorithm procedure via S-plus software. Maximum posterior probability is used to classified trials to each component.

Results: The two approaches shown that the covariates effects and variability of random effects from the models easily explained heterogeneity between trials. Results of numerical

examples are presented in topic 6 and 7. The GLMM is superior to the GLM in some aspects. The GLMM gives further heterogeneity information from random treatment effects. In addition, the approach provides component (or subgroup)-specific treatment effect and trial classification according to the optimal components. This is very useful in further explaining the heterogeneity that might be beyond the effects found in the model.

Conclusions: The GLMM approach provides more information for explaining heterogeneity effect in meta-analyses involving cluster randomized trials. However, care should be taken when interpreting the covariates effects of the model because inference on these effects obtained from a discrete mixing distribution have not been ruled out. Nevertheless, the GLMM would be much more efficient when it is applied to large meta-analyses.

บทคัดย่อ

ความเป็นมา : วิธีการวิเคราะห์หน่วยเดียวกับทางสถิติที่ใช้ใน meta–analysis ส่วนใหญ่เป็นวิธีการที่ใช้กับข้อมูล ซึ่งมาจากหน่วยศึกษาที่เป็นหน่วยเดียวกับหน่วยสุ่ม meta-analysis ที่เกี่ยวข้องกับผลลัพธ์จากการทดลองที่มีกลุ่ม (cluster) ของหน่วยศึกษา (individual units) เป็นหน่วยสุ่มรับสิ่งทดลองอาจทำให้เกิดความแตกต่าง (heterogeneity) ที่นอกเหนือจาก meta-analysis ที่มีหน่วยศึกษาเป็นหน่วยเดียวกับหน่วยสุ่มรับส่งทดลอง วิธีการทางสถิติที่เหมาะสมกับ meta-analysis ที่เกี่ยวข้องผลลัพธ์จากการทดลองที่มี clusters ของหน่วยศึกษาเป็นหน่วยสุ่มรับสิ่งทดลอง จะต้องพิจารณาอิทธิพลที่เกิดจาก clusters ด้วย

ได้มีการเสนอ General linear mixed model (GLM) มาวิเคราะห์ข้อมูลของผลลัพธ์สิ่งทดลองที่วัดออกมาเป็นข้อมูล binary โดยใช้ค่า log-relative measure เป็นตัวแปรตามของโมเดล การประมาณค่าพารามิเตอร์จะขึ้นอยู่กับการแจกแจงแบบปกติของ random effects Generalize linear mixed model (GLMM) เป็น model ที่ไม่ต้องการรูปแบบการแจกแจงของ random effects ในการประมาณพารามิเตอร์ GLMM อาจเป็นอีกวิธีหนึ่งที่นำมาใช้ได้ GLM และ GLMM สามารถให้ข้อมูลอิทธิพลของ covariates ทั้งที่ระดับสิ่งทดลองและระดับหน่วยศึกษา ดังนั้นการศึกษาความสามารถของวิธี GLM และ GLMM ในการวิเคราะห์ข้อมูลของ meta-analysis ที่เกี่ยวข้องกับผลลัพธ์จากการทดลองที่มีกลุ่มของหน่วยศึกษาเป็นหน่วยสุ่มรับส่งทดลอง จึงเป็นสิ่งที่น่าสนใจ

วัตถุประสงค์ : เพื่อศึกษาความสามารถของ GLM และ GLMM ในการตรวจสอบและอธิบายความแตกต่างใน meta-analysis ที่เกี่ยวข้องกับผลลัพธ์จากการทดลองที่มีกลุ่มของหน่วยศึกษาเป็นหน่วยสุ่มรับส่งทดลอง ซึ่งวัดประสิทธิภาพสิ่งทดลองเป็น binary data

วิธีการศึกษา : ในการศึกษาครั้งนี้ ได้ทำการประเมินความสามารถของ GLM และ GLMM ในประเด็นเกี่ยวกับ วิธีการ ผลลัพธ์ ที่ได้จากแต่ละวิธี ข้อดีและข้อจำกัดของแต่ละวิธี การใช้ GLM และ GLMM ในการวิเคราะห์ข้อมูลจริงจาก meta-analysis ดังกล่าว 3 เรื่อง ได้แก่เรื่องแรก meta-analysis ที่ศึกษาเพื่อหาความสัมพันธ์ระหว่างวิตามินเอ และการตายในเด็กอายุ 6 ถึง 72 เดือน โดยรวบรวมข้อมูลจาก 8 การทดลองที่ศึกษาในชุมชนของประเทศกำลังพัฒนา และในทุกการทดลองไม่มีการสุ่มเด็กรับวิตามินเอ เรื่องที่สองเป็น meta-analysis ของข้อมูลการทดลอง 8 เรื่อง ที่ศึกษาประสิทธิภาพของ memmographic screening ในการลดการตายของมะเร็งเต้านม เรื่องที่สามเป็นเรื่อง meta-analysis ของข้อมูลการทดลอง 14 เรื่อง ที่ศึกษาประสิทธิภาพของ multiple risk factor interventions ในการลดปัจจัยเสี่ยงของโรคหัวใจโคโรนารี่ โดยมีอัตราชุกของการสูบบุหรี่เป็นผลลัพธ์ที่ใช้วิเคราะห์ การวิเคราะห์ด้วย GLM ใน meta-analysis แต่ละเรื่อง ค่าสังเกต log-relative risk ของแต่ละการทดลองจะนำมาใช้เป็นตัวแปรตามแบบต่อเนื่องในโมเดล ลักษณะของหน่วยสุ่มรับสิ่งทดลองนำมาศึกษาเป็นตัวแปร covariates ที่แยกออกเป็น 2 กลุ่มตามลักษณะของหน่วยสุ่มรับสิ่งทดลอง ซึ่งอาจจะเป็นแต่ละหน่วยศึกษา (individual) หรือ กลุ่มของ

หน่วยศึกษา (cluster) การประมาณพารามิเตอร์ของโมเดลใช้ restricted maximum likelihood ภายใต้ข้อสมมุติของการแจกแจงแบบปกติของ random effects โดยวิเคราะห์ด้วยโปรแกรม ML win

สำหรับการวิเคราะห์ข้อมูลดังกล่าวด้วย GLMM ค่าสังเกตจำนวนของผลลัพธ์ที่เกิดขึ้นในแต่ละกลุ่มของสิ่งทดลองของการทดลองแต่ละเรื่อง คือ ข้อมูลของตัวแปรตาม โมเดล canonical link function ของค่าสังเกตค่าเฉลี่ยสัดส่วนจะสร้างจากโมเดลตัวประมาณค่าเชิงเส้นตรงที่มีสิ่งทดลองและแบบการสุ่ม (randomization design) เป็นตัวแปร covariates ในการวิเคราะห์ของ GLMM สามารถกำหนดให้อิทธิพลของสิ่งทดลองเป็นแบบสุ่มได้ การประมาณค่าพารามิเตอร์ของโมเดลใช้วิธี non-parametric maximum likelihood ภายใต้การแจกแจงของ random effects แบบไม่ต่อเนื่อง ที่แยกออกเป็น k components ซึ่งทำการวิเคราะห์โดยใช้ EM-algorithm ด้วยโปรแกรม S-plus Maximum posterior probability ที่ได้จากการวิเคราะห์นำมาใช้ในการจัดกลุ่มการทดลองให้กับแต่ละ components

ผลลัพธ์ของการศึกษา : จากการวิเคราะห์ด้วย GLM และ GLMM แสดงให้เห็นว่า อิทธิพลของ covariates และความแตกต่างของ random effects สามารถใช้อธิบายความแตกต่างระหว่างการทดลองได้อย่างง่ายดาย ผลลัพธ์ในเชิงตัวเลขที่ได้จากการวิเคราะห์ในข้อมูลตัวอย่าง meta-analysis ได้แสดงไว้ในหัวข้อที่ 6 และ 7 ของรายงานฉบับนี้ และพบว่าวิธีของ GLMM ดีกว่า GLM ในบางประเด็นซึ่งได้แก่ ความสามารถในการวิเคราะห์อิทธิพลของสิ่งทดลองเชิงสุ่ม ซึ่งนำมาใช้อธิบายข้อมูลความแตกต่างระหว่างการทดลองได้ นอกจากนี้ GLMM ยังให้ข้อมูลเกี่ยวกับอิทธิพลของสิ่งทดลองเฉพาะในแต่ละ component ซึ่งนำมาใช้ในการแยกการทดลองเป็นกลุ่มๆ ตามจำนวนของ optimal components ข้อมูลเหล่านี้เป็นประโยชน์อย่างยิ่งในการอธิบายความแตกต่างระหว่างการทดลองที่เกิดขึ้น ซึ่งเป็นอิทธิพลที่นอกเหนือขอบเขตของข้อมูลที่อธิบายได้จากโมเดล

สรุปผลการศึกษา : GLMM เป็นวิธีการวิเคราะห์ที่สามารถใช้ข้อมูลเพื่ออธิบายความแตกต่างที่เกิดขึ้นใน meta-analysis ที่เกี่ยวข้องกับการทดลองที่มีกลุ่มของหน่วยศึกษาเป็นหน่วยสุ่มรับสิ่งทดลอง อย่างไรก็ตามควรต้องระมัดระวังในการแปลผลอิทธิพลของ covariates เนื่องจากยังไม่มีหลักฐานการตรวจสอบความถูกต้องของการอนุมานอิทธิพลดังกล่าวที่วิเคราะห์ได้จาก discrete mixing distribution แม้จะด้วยเหตุผลดังกล่าว แต่พบว่าการใช้ GLMM ใน meta-analysis ที่รวบรวมจำนวนการทดลองมาศึกษาเป็นจำนวนมากจะให้ผลลัพธ์ที่มีประสิทธิภาพดีขึ้น

Executive Summary

Title: Combining Cluster and Individual Randomized Trials in Meta-analysis

Summary

In this study, the main aim is to explore and compare two potential non-Bayesian approaches of GLM and GLMM to identify and explain treatment heterogeneity in meta-analyses involving cluster randomized trials measured in binary outcome. General concepts on designs and analysis of cluster randomized trials are illustrated in topic 2. A review of current practice on meta-analysis related to cluster randomized trials is discussed in topic 3. Three meta-analyses (1-3) on different situations according to number of trials and randomization designs, which are described in topic 4, are selected from the published literature. They are used to illustrate application of individual approaches. A review of some essential issues in meta-analysis and simple conventional approaches are discussed in topic 5.

Exploration starts with the GLM approach, a linear regression mixed model. Details of the approach are presented in topic 6. The model allows potential factors to be added both in continuous and discrete form. For the three examples, log-relative risks of individual trials are used as a continuous response variable. The covariate is a randomization design, which is a binary variable for the two examples(2, 3). Estimation of model parameters is based on the assumption of normal distributions of random effects. The parameters are estimated by REML via the RIGLS algorithms. Because the number of trials in the examples are small, standard errors of the estimates provided by the REML based on asymptotic properties, especially for the variance τ^2, may be unreliable. Therefore, the parametric bootstrap estimation is used to calculate confidence intervals for all the parameters. The GLM produces estimates of an adjusted overall treatment effect, covariate effects and variance of random effects to explain heterogeneity. Applications of the GLM to the three examples are implemented using the MLwiN software. This approach is logically appropriate to the meta-analyses related to cluster randomized trials.

Further investigation is performed in the GLMM. The approach is described in topic 7. This is a regression mixture model allowing the inclusion of some covariates of trial level and subject level. The GLMM is utilized in a two-levels mixed poisson regression models, applied to the three examples. Observed numbers of events from individual trials are used as a response variable. For the investigation of random treatment effects, the two treatment groups are treated as a binary covariate in the model. The other covariate is randomization design, which is also a binary variable. The random treatment effect is obtained from an interaction term between intercept and treatment variable. The NPML estimates of the parameters are obtained from the discrete mixing distribution of the random effects via the EM-algorithm. The optimal number of components is selected using the BIC- criterion. The results of component-specific mean treatment effect and component weight reflect the heterogeneity due to random treatment effects. When the optimal number of component is more than one (K>1), further trial classification is performed using the posterior probability. Applications of the GLMM to the three examples are implemented using the S-plus software.

These different approaches are then evaluated in topic 8 for their potential in terms of application to the meta-analyses related to cluster randomized trials. Items to be evaluated are methodological issues, heterogeneity information provided, model complexity, interpretation of the results, strengths and limitations and numerical results of the three examples.

In conclusion, the GLMM are comparable in term of methodology aspects. Their aim of analysis is appropriate to investigate heterogeneity effect in the meta-analyses. They are both attainable approaches that provide results to be used to explain all dimensions of sources

of the heterogeneity in the meta-analyses. The GLMM is superior to the GLM in some aspects. The GLMM gives further heterogeneity information from random treatment effects. In addition, the approach provides component (or subgroup)-specific treatment effect and trial classification according to the optimal components. This is very useful in further explaining the heterogeneity that might be beyond the effects found in the model. However, some limitations of the two approaches have to be considered. Meta-analysts should have literate statistical modelling to perform the approaches in meta-analyses data. The results obtained from the GLM may be unreliable if the normal distributed assumption of random effects is misspecified. Also, the GLMM may have some difficulty in generalizability of the estimated treatment effect as the unsolved problem in asymptotic inference from nonparametric approach. This issue should be kept in mind and carefully considered when interpreting the treatment effect. According to the considerations, GLMM is a preferable choice.

Suggestions

Although the proposed GLMM in this study is investigated under the scope of meta-analyses involving cluster randomized trials, they can also be applied to other meta-analyses that have some potential covariates available at trial level and subject level. In terms of software, the S-plus is used in this study. For this approach several software, such as GLIM and STATA, are also available.

Although heterogeneity is one of the main issues in meta-analysis, interpretation of treatment effects is also sometimes required. Therefore, researches need to be done to solve the problem in the interpretation of estimated treatment effect from the NPML estimator of mixing distribution.

Content

Topic1: Introduction

1.1 Background

The randomized controlled trial is a well-established method used to evaluate the effectiveness of treatments in health care. Most of the trials involve treatment allocation to individual subjects, such as patients, school children and villagers. Such trials are called *individually randomized trials* (IRTs). The effect of treatments is measured at individual subjects, which is the unit of randomization. The individual subjects are, therefore, assumed to be independent in terms of their responses. Variation of the responses comes from variation among the individual subjects of each treatment group.

Throughout the 1980s and 1990s, different trial designs have been increasingly used to evaluate treatment effectiveness. An important modification of the IRT is the *cluster randomized trials* (CRTs). The cluster randomized trials are particularly relevant in field trials on tropical diseases in developing countries such as in Thailand. In these trials, treatments are randomly assigned to clusters (or groups) of individuals. Examples of the units of treatment allocation are villages, schools, work sites, general practices, and hospitals.

An example of the cluster randomized trial is the WHO trial (4) conducted to evaluate a new antenatal care programme compared to the standard antenatal care programme in pregnant women. Fifty-three antenatal care clinics are stratified by countries and clinic sizes. The clinics are randomly allocated to each of the antenatal care programmes. Therefore, each clinic would have only one antenatal care programme. Such that would make it more convenient for investigators to manage the trial and also to avoid pregnant women receiving more than one antenatal care programme. This is why cluster randomization is used instead of individual randomization.

The treatment effects of cluster randomized trials are measured for individual subjects nested within the clusters. Hence, the responses of individuals in the same cluster cannot be regarded as independent. The responses in the same cluster tend to be more similar, compared to those obtained from subjects in different clusters. For example, newborn babies in the same nursery ward would likely have similar infection rate as compared to the babies in different nursery wards. This is because the newborn babies within the same nursery would be exposed to the same temperature as compared to those in different nurseries. So, variations of the responses in a cluster randomized trial are measured from two sources, i.e. within a cluster and between clusters.

The similarity of individual responses within the same cluster is reflected in a measurable *intracluster correlation coefficient(ICC)*. The ICC is a ratio of the variation of between clusters to the overall variation, which is the addition of between and within cluster variations. The ICC is used to calculate *clustering effect*, which is the specific characteristic of cluster randomized trials. The ICC introduces one or more extra sources of random variation that must be reflected in the sample size determination and data analysis. The clustering effect is generally known as *design effect*. To account for the clustering effect in data analysis, an analysis at cluster or individual level can be done depending on the research question.

The randomization designs of cluster randomized trials used in health care evaluation are commonly classified into three designs, *completely randomized, matched-pair randomized and stratified randomized.* The choice of the designs depends on nature and available number of clusters. Rationale for each design is given in the next section. The units of randomization may vary even though the trials are conducted to evaluate similar treatments. For example the eight cluster randomized trials included in the meta-analysis of Vitamin A supplementation and child mortality(1) differ in the units of randomization. The

randomization units are clusters of children, households, villages, areas, sub-districts and wards. This issue may raise heterogeneity in the meta-analysis.

A large amount of the literature on randomized controlled trials has been published in the last two decades. There is, however, some controversy on conclusions found in the trials to assess similar or the same treatment effect. Furthermore, some trials were done in very small sample sizes. The conclusions on the treatment effect of such trials are, thus, questionable.

Meta-analysis is known as the statistical method used to gather and combine information from many related trials(5). The method aims to compare and potentially combine estimates of treatment effect across trials(6). Meta-analysis may provide a clear overall picture when single trials may appear inconsistent with regards to degree or direction of the treatment effects of interest. By including several trials, the results can be more precise and may also allow investigation and identification of variability of treatment effects across trials (7, 8).

Since the mid-1980s meta-analysis has become an important part of health care research not only primarily for randomized controlled trials but also for observational epidemiological studies(9). Continuing activities on meta-analysis have increased steadily since the past decade. When the PubMed electronic database of the National Library of Medicine is searched using 'meta analysis' as a key word, only 6 papers are found in the year 1980. The number increased to 323 papers in 1990. A big increase of up to 1,175 papers is seen ten years later in 2000. One reason of such an increase is the awareness of a demand to obtain not just evidence but reliable evidence from all the relevant studies(10). The reliable evidence is used to justify any decision in health care activity, such as whether to give a new therapeutic treatment to diabetic patients and whether mammographic screening should be kept or abandoned, etc.

Meta-analysis can yield numerical statistics for overall treatment effects of interest, such as relative risk, odds ratio, mean differences, and confidence intervals via several existing estimation approaches both in Frequentist and Bayesian perspectives(11, 12). These statistics have been most commonly applied to the results of individually randomized trials, which is the most common type of randomized controlled trials.

The variability of treatment effects between the trials is likely, because of the random chance, known as sampling error. If estimates of treatment effect vary among trials beyond that expected by chance alone, it is generally known as '*heterogeneity*' in meta-analyses. Potential sources of heterogeneity may be from some biases due to different trial and subject characteristics between trials, and unobserved random effects(13). It would be clearly remarkable if all the trials being meta-analysed yield the same treatment effect. However, in practice it is hardly possible for such trials to be included in the meta-analyses. So, heterogeneity is an important issue that must be explored and identified when a meta-analysis is applied(6, 14-17).

Meta-analyses involving cluster randomized trials are increasingly found in many health care publications. These meta-analyses may lead to additional sources of heterogeneity beyond those elevated by meta-analyses involving only individually randomized trials. Papers of cluster randomized trials may be different in eligibility criteria on both the cluster and individual level, in randomized designs and units of randomization. In addition, they may present results from different levels of units of analysis, cluster and individual. Simple conventional methods, both in the area of fixed effect and random effects models, ignoring this heterogeneity may result in incorrect inferences for the treatment effects. However, the empirical evidence from a review of 25 published meta-analyses related to cluster randomized trials shows that 15 of the meta-analyses used simple conventional methods of the fixed effect model as method of analysis. Detail of the review is illustrated in Topic 3.

The appropriate statistical analysis to these meta-analyses must take into account potential heterogeneity in the cluster randomized trials.

A substantial amount of literature covering statistical methodologies used in meta analyses can now be found. Most of them, however, assume individual subjects as units of randomization. Therefore, there may remain some questions that need to be investigated in the area of meta analyses related to the inclusion of cluster randomized trials.

1.2 Aims and objectives

In this study some potential non-Bayesian approaches are explored to identify and explain heterogeneity in meta-analyses involving cluster randomized trials comparing two treatment groups measured in binary outcome.

The specific objectives are:

- to investigate the potential of general linear mixed model (GLM) and generalized linear mixed model(GLMM) with nonparametric maximum likelihood estimator (NPMLE) to identify and explain heterogeneity in the meta-analyses;
- to compare methodological aspects, results provided, strengths and limitations of these approaches to the common approaches of simple conventional methods, and
- to propose appropriate approaches to apply to the meta-analyses related to cluster randomized trials.

1.3 Sequences of further topics

This study covers 8 topics. Literature review is presented in topics two, three and five. Topic 2 describes general concepts of design and analysis in cluster randomized trials. Topic 3 presents a review of meta-analyses involving cluster randomized trials and discusses current practice in the meta-analyses. Topic 4 introduces examples of three published meta-analyses in binary outcome to be used in the study. Topic 5 discusses important elements in the analytical approach of meta-analysis. This topic also discusses simple conventional approaches. Methodology, results and discussion are presented together in topics six, seven and eight. Topic 6 discusses the GLM and illustrates its application to the three examples. Topic 7 discusses the GLMM and illustrates its application to the three examples. Topic 8 presents a comparison of the two approaches in terms of their methodology issues, results provided and interpretation of results. The topic also discusses strengths and limitations of individual approaches. Finally, the topic presents the proposed approaches to be used in meta-analyses related to cluster randomized trials under the discussed situations in the topic.

Topic 2: General concepts of design and analysis in cluster randomized trials

2.1 Introduction

Most randomized controlled trials involve treatment allocation to individual subjects. However, there are many situations in health care areas where such allocation is not desirable or even possible. Instead, clusters or groups of individuals may be randomized to a treatment or control group. Such trials are often called cluster-randomized trials. This term will be used throughout this study. Other terms are group-randomized trials, community intervention trials, etc. The units of treatment allocation of cluster-randomized trials vary. They range from relatively small clusters like households or families to relatively large ones like entire villages or, communities (18).

This topic describes the general concepts of a cluster randomized trial. Section 2.2 discusses the rationale for performing cluster randomization. Section 2.3 presents variation in cluster randomized trials and clustering effect. Section 2.4 describes the common alternative designs of cluster randomized trials applied to health research. Section 2.5 discusses the analysis issue. Section 2.6 discusses the problems found in published papers on cluster randomized trials.

2.2 Rationale for cluster randomized trials

Over the past two decades cluster-randomized trials have been markedly increasingly used to study effectiveness of health intervention. There are reasons for using clusters as the units of treatment allocation (19-24).

The first reason is specific to the circumstance of infectious disease (24). Some treatments are aimed primarily at disease transmission to the other persons come in contact with the subjects receiving the treatments. The treatment effect is also expected on decreasing the susceptibility of subjects receiving such treatment. An example is the randomized community trial of improved control sexually-transmitted diseases (STDs) conducted for AIDS prevention in Uganda (25). The STDs treatment was expected to block onward HIV-infected subjects from passing the infection to their partners. This situation cannot be answered by individually randomized trial where the effect of treatment is measured from the subjects who received treatment. Alternatively, a cluster-randomized trial allows the overall treatment effect on both infectious and susceptibility to be captured at community level (24).

The second reason is that some treatments may be maximized if the treatments are received by a large proportion of the population (24). An example is the trial on providing impregnated bednets for malaria control in Africa (26). The investigators expected that, with the high usage of nets in a village, the overall level of transmission of malaria might be reduced through the mass killing of the mosquito vectors, which in turn would also protect non-net users from infection.

The third reason is when treatment implementation cannot be directed to individual subjects. An example is a community trial on the impact of improved sexually-transmitted diseases (STD) on HIV epidemic in rural Tanzania (27). The STD treatment intervention involves the provision of improved services at health facilities. The services offered by each facility are available to the entire population. This necessitates the randomization of whole communities, rather than individuals.

Other reasons include administrative convenience, political aspect and avoiding treatment contamination (19, 20). In some instances it may not be convenient from an administrative or political viewpoint to allocate individuals of the same cluster to different intervention groups (23). An example is a WHO trial carried out to evaluate whether a new programme of antenatal care, which only includes item of care of proven effectiveness, has similar outcomes to current standard care. Fifty-three antenatal care clinics were randomized

to each of the two intervention groups rather than randomizing women within the same clinic to different intervention groups.

The need to avoid treatment contamination is also a common reason for choosing the cluster randomization design (20, 21). The WHO antenatal care trial is one example of this. If the women within the same clinic were randomly allocated to different intervention groups, the women allocated to the new intervention group may adopt the antenatal care strategies of the standard group. This is because the women of the intervention group have less antenatal care visits than the standard group.

2.3 Variation and clustering effect

In the individually randomized trial, different treatments are randomly allocated to individual subjects. Analysis is also performed for the individual subjects. These individual subjects are then assumed to be independent in terms of their responses. Therefore, variation of the treatment outcome comes from variation among the individual subjects due to treatment group.

In contrast, for the cluster randomized trial, the unit of randomization is the cluster of individual subjects, which is a higher aggregated level than that of the individual randomized trial. The responses of the subjects within the same cluster may be more similar than those of other clusters. For example, children in the same class receive the same teaching pattern from the same teacher. They also share the same discussion experience in their class. This combination increases the likelihood that the children will respond similarly in trial where classes are the randomization units. In such a situation, the outcome of interest is measured for individual subjects nested within individual clusters. So, variation of the outcome comes from two sources; within a cluster and between clusters.

The intra-cluster correlation coefficient (ICC; ρ) is a measure of within-cluster similarity or homogeneity. It is defined (28) as the proportion of the total variation of the response accounting for differences among the clusters. The ICC is expressed as

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \tag{2.1}$$

where σ_b^2 is the between-cluster variance and σ_w^2 is the within-cluster variance.

By applying the analysis of variance, σ_b^2 and σ_w^2 can be estimated (29) by

$$\hat{\sigma}_b^2 = \frac{MSB - MSW}{n_0} \tag{2.2}$$

$$\hat{\sigma}_w^2 = MSW \tag{2.3}$$

where MSB is the between-cluster mean square and MSW is the within-cluster mean square, n_0 is the average cluster size obtained by

$$n_0 = [\frac{1}{J-1}][N - \frac{\sum n_j^2}{N}] \tag{2.4}$$

where J is the number of clusters, N is the total number of individuals, and n_j is the number of individuals in the j^{th} cluster. Equation (2.2) and (2.3) can substitute into equation (2.1) and then $\hat{\rho}$ is given by

$$\hat{\rho} = \frac{MSB - MSW}{MSB + (n_0 - 1)MSW} \tag{2.5}$$

Commonly, the ICC is a positive value with maximum at 1. Negative values of the ICC are usually set to zero because values of ICC less than zero are generally considered

implausible in the context of cluster randomized trial(19). The values of ICC tend to be larger in smaller clusters but with a non-linear relationship (30, 31).

The design effect (Deff) of cluster randomization is defined (28) as a ratio of the variance of the estimated outcome under the cluster randomization (σ_C^2) to the variance of that under simple random sampling (σ_{SRS}^2) with the same number of individuals (or sample size),

$$\text{Deff} = \frac{\sigma_C^2}{\sigma_{SRS}^2} \qquad (2.6).$$

The design effect is sometimes called clustering effect in cluster randomized trials. It can be interpreted as the multiplying factor of the number of subjects in cluster randomized trials, compared with the number of subjects in individually randomized trials required to get the same power. The sample size obtained by the standard method should be multiplied by the design effect to compensate for clustering effect in a cluster randomized trial. The analysis of cluster randomized trial also needs to account for the design effect so that the inference will be identified as valid (23, 32). The design effect can also be obtained by,

$$\text{Deff} = 1 + (n-1)\,\rho \qquad (2.7)$$

where n is the average cluster size for balanced cluster sizes trials (28). But when the trials vary in cluster size, n is recommended to replace with n_0 from the formula (2.4) (23).

The design effect can be large for large clusters even with small ICC. It can be unity if there is no between-cluster variation ($\sigma_b^2 = 0$) because ρ is zero.

2.4 Strategies for randomization in cluster trials

When performing cluster randomized trials, the decision to apply an appropriate randomization design depends on a number of factors, e.g. number of potential clusters in the study and baseline characteristics. In the completely randomized design, assigning different treatments to clusters is done without pre-matching or stratification by potential factors related to outcome. It is most appropriate when there are many clusters available to be randomized(33). When there are small numbers of clusters, it may yield an unbalance between treatment groups with respect to baseline characteristics. The completely randomized design is not recommended under these circumstances (23, 29).

An example is a randomized control community trial done to evaluate the impact of vitamin A supplementation on child mortality in northern Sumatra, Indonesia (34). 450 villages were randomly assigned to either participate in a vitamin A supplementation scheme (n=229) or serve for one year as a control (n=221). Child mortality rate at one year of follow-up was the measure of the effect of vitamin A supplementation.

Matched-pair randomization is a design where clusters are matched in pairs according to baseline characteristics such as cluster size, demographic characteristics or other potential factors associated with the outcome. From every pair, one cluster is then assigned to each treatment group at random. The main advantage of this design is that it gives very close balancing of important baseline risk factors (20). However, there are also disadvantages(19, 20, 23, 29, 35-38). When there are few clusters, it may be difficult to get close matches on all potential risk factors. In addition when a particular cluster of any pair drops out, the pair has to be deleted from the study. This will lead to a decrease in the study degrees of freedom. Furthermore, one cannot estimate all the variance components since the between-cluster variation cannot usually be estimated within pairs. This is because each cluster within a pair receives a different treatment. In addition, between-clusters variation within the same treatment group cannot be estimated due to confounding with differences between pairs.

An example is a randomized community trial conducted to assess impact of improved treatment of sexually-transmitted disease (STD) on HIV infection in rural Tranzania (39). Twelve large communities were matched on area, type and prior STD rate into six pairs. One of each pair is then randomized to the intervention group. Two years of HIV incidence is the outcome measure.

Stratification is an extension of the matched-pair design, in which several clusters within each stratum are randomly assigned to each treatment group. This design should reduce the probability of large imbalances on important prognostic factors (19, 20, 23). Stratified design provides some advantages. Since there is more than one cluster within each treatment-stratum combination, it is possible to estimate between-cluster variation. This is because the clustering effect can be separated from both the treatment effect and the stratum effect. Stratification by cluster size is considered desirable not only to accomplish balancing the number of individuals in each treatment group, but also the cluster size may be a marker for within-cluster dynamics that is predictive for outcome. When a large number of clusters relative to the number of confounding factors exist, it is easier to construct meaningful strata under the stratified design as compared to the pair-matched design.

An example is a WHO trial to evaluate a new antenatal care programme (4). The trial was done to evaluate whether a new programme of antenatal care, which only includes items of care of proven effectiveness, has similar outcomes to the current standard care. Fifty-three antenatal care clinics were randomized to each programme after stratification by country (4 different countries) and clinic size (small, medium and large). Stratification according to country is expected to provide some control over confounding by country. Clinic size is an additional stratification factor that is expected to be a marker for a range of baseline factors. The main outcomes are: low birth weight, pre-eclampsia/eclampsia, severe postpartum anemia and treated urinary-tract infection.

2.5 Analysis issues

As clusters are the units of randomization rather than individuals, the clustering effect must be taken into account in the analysis. The simplest approach is to use the cluster as the unit of analysis. Then, a summary statistic for all individuals within each cluster is the outcome variable, e.g. hospital mean length of stay, area utilization rate, proportion of smokers. Standard statistical methods can be used to compare the cluster responses between different treatments, and thus if it is necessary, cluster-level baseline risk factors, such as cluster sizes or urban/rural location, can be adjusted for. Obviously, it is appropriate to do cluster level analyses when the inference of the trial focuses directly on the randomization unit as a whole rather than on the individual subjects. However, one has to keep in mind that a large number of clusters is needed. Otherwise, it may lack power of study (19).

Another appropriate approach is to use the individual as the unit of analysis, adjusting for the dependency among individual responses within the same cluster. To do valid individual level analyses, a large number of clusters per treatment group are required to adequately account for the clustering effect(19).

Substantial literature on basic approaches without considering covariates and advanced approaches of modelling analysed at individual subjects have been proposed to use in the cluster randomized trials (23). Mixed effect models that account for the variability between clusters have appeared frequently in the literature on cluster randomized trials. Most of these are based on assumption of normal distribution of random effects. Furthermore, Bayesian approaches as an extension to the hierarchical modelling are becoming increasingly used for cluster randomized trials (40). This development took place because of some deficiencies found in the former approaches based upon the normality assumption of the

random effects. Another reason might be because of the availability of advances in computational methods.

2.6 Problems in published papers of cluster randomized trials

Three papers have specifically reviewed the methodology of design and analysis aspects in the published literature of the cluster randomized trials in areas of health care(23, 32, 33, 41). The problems found seem to persist from 1979 to 1996. Most studies did not recognise the clustering effect in the sample size estimation. Thus the sample size may not be big enough to give reliable estimate of treatment effects. Studies are often designed in small number of clusters. Sometimes only one cluster is randomized to each of two or more treatment groups (23, 41). Studies often provide the results at an individual level without adjusting for clustering effect. This potentially leads to spurious statistical significance and too narrow confidence interval of the treatment effect of interest. Few studies explicitly published clustering effect information in their trials. This might possibly lead to difficulties to the statistical considerations for future investigators in planning the total number of clusters to be randomized.

These problems need to be considered when including cluster randomized trials from the published literature to meta-analyses.

Topic 3: Review of meta-analyses involving cluster randomized trials

3.1 Introduction

Meta-analysis of the trial results is now a common tool used in health care research. It could yield more reliable conclusion of treatment effects than that obtained from individual trials alone. There is now substantial literature to be found covering the statistical methodology used in meta-analyses. Most of them relate to meta-analyses of trials, which randomize individual subjects. Cluster randomized trials have received less research attention in the literature on meta-analysis. Thus, to obtain empirical evidence of the current situation, a review is conducted in published meta-analyses involving cluster randomized trials. The objectives are to describe statistical approaches for handling heterogeneity and estimation of treatment effects in the meta-analyses involving cluster randomized trials.

Section 3.2 describes the methods used for the review, including strategies to search and identify trials from the published literature. The reviewing procedure for each paper is also provided in this section. Section 3.3 describes results obtained from the review. Section 3.4 presents comprehension of the results and makes a conclusion from the review.

3.2 Methods

3.2.1 Study search and identification

Electronic search was performed for reports in English on meta-analyses involving cluster randomization trials. We were aware of difficulty in searching the reports related to cluster randomization trials. It was because using the keywords of 'cluster randomization' might not be able to identify some of the meta-analyses involving such trials. Therefore other keywords related to 'cluster randomization' were also combined with the keywords of meta-analyses. The search keywords were presented in Table 3.1. The following electronic databases were used: Medline, Health Star, Embase, SCIsearch and the Cochrane Library. The SCIsearch database was used to identify further references that cited the relevant papers. The search was done from the first year of each electronic database to 2000.

Once a meta-analysis was identified, papers on the relevant cluster randomization trials included were also requested.

Table 3.1 Keywords used for electronic databases searching

1 meta-analysis	14 (10) and (13)
2 randomized controlled trials	15 (10) and (4)
3 randomised controlled trials	16 trials
4 (2) or (3)	17 intervention trials
5 (1) and (4)	18 (16) or (17)
6 cluster	19 (10) and (18)
7 group	20 cluster effect
8 community	21 design effect
9 field	22 inflation factor
10 (6) or (7) or (8) or (9)	23 intracluster correlation
11 randomization	24 clustering
12 randomisation	25 (14) or (15) or (19)
13 (11) or (12)	26 (1) and (25)

3.2.2 Review process

Each cluster randomization trial was reviewed with respect to designs of randomization and adjustment for clustering effect in the analysis. Each meta-analysis was then reviewed with respect to number of trials included, particularly the number of cluster randomization trials, types of intervention of interest, outcome measure, methods to obtain an overall treatment effect and heterogeneity consideration regarding the inclusion of cluster randomization trials. The interventions of interest were classified into three main types, educational, health care and screening. The educational intervention was referred to the interventions related to health promotion or non-therapeutic treatments, e.g. mass media, group behavior therapy, etc. The health care intervention was referred to the interventions related to therapeutic or preventive treatments, e.g. routine antenatal care, vitamin A supplementation, etc. The screening was referred to the interventions related to investigation of disease in general people, e.g. memmographic screening, etc.

3.3 Results

The search identified 25 eligible meta-analysis reports published between January 1990 and 2000. Sixteen reports were from the Cochrane Library, and two were from the British Medical Journal. Each of the remaining seven was from the American Journal of Public Health, American Journal of Tropical Medicine and Hygiene, Bulletin of the World Health Organization, International Journal of STD&AIDS, Journal of American Medical Association, Journal of the National Cancer Institute Monographs and The Medical Journal of Australia, respectively.

Table 3.2 presents types of intervention studied and trials included in each of the 25 meta-analyses. Health care intervention was the majority, which accounted for 64.0 per cent (16/25) of the meta-analyses. A total of 89 cluster randomization trials and 297 individually randomization trials were included in these 25 meta-analyses. A mean number of 15 trials was found for individual meta-analyses ranging from 2 to 41. For the cluster randomization trials included, a mean of 4 ranging from 1 to 17 was found. There were 15 meta-analyses that included more than one cluster randomization trials. The randomized units of cluster randomization trials within the same meta-analysis were mostly different. For example, in a meta-analysis on mass media interventions to prevent smoking among children (42), the three included cluster randomization trials had area, school and community as a randomized unit, respectively. Moreover, eligibility criteria at both cluster and individual level of the trials included in the same meta-analysis were quite different. These differences among the cluster randomization trials might lead to extra sources of heterogeneity beyond those already existed in meta-analysis including only individually randomization trials. Consequently, they might raise more difficulties in methodologic issues.

From the 89 cluster randomization trials, 83 original papers could be reviewed. In two of the remaining six cluster randomization trials, the required information was extracted from the meta-analyses in which they were included. One of them was an unpublished paper, and the other was written in Russian. The remaining four cluster randomization trials could not be accessed as they were referenced incorrectly. We attempted to search for these four trial papers but did not succeed to access the correct papers. Consequently, a total of 85 cluster randomization trials could be reviewed. References of the trials reviewed are presented in Appendix 1. The following results were thus based on the accessible papers.

Twenty-two meta-analyses had a binary endpoint as the primary outcome. One meta-analysis had a binary and a continuous endpoints as the co-primary outcomes. Fifteen meta-analyses reported simple conventional methods of fixed effect model as methods of analysis. They treated the cluster randomization trial results as individually randomization trial results.

Six meta-analyses did not incorporate the cluster randomization trial results in the quantitative synthesis. They described the results of cluster randomization trials separately. Three meta-analyses reported the synthesis methods that account for clustering effect. One was unclear, as it did not report the synthesis method.

Details of randomized design and unit of analysis for each cluster randomization trials included in each meta-analysis and the combining methods were presented in Table 3.3. Here the last three columns were considered together. In the group of fifteen meta-analyses that reported simple conventional methods in the quantitative synthesis, two meta-analyses(43, 44) likely provided reasonable evidence because the results of cluster randomization trials included were analysed at individual unit adjusted for clustering effect. Nine of the fifteen meta-analyses included cluster randomization trials with a mixture of different randomized designs. They were completely randomized, matched-pair randomized and stratified randomized. The cluster randomization trials included in the nine meta-analyses also had a mixture of different units of analysis, some at cluster level and some at individual level. These mixtures certainly raised additional heterogeneity in the meta-analyses and needed to be considered in the synthesis procedures. However, none of these meta-analyses reported any concern on the heterogeneity that might be due to cluster randomization trials.

For the six meta-analyses that did not incorporate the results of cluster randomization trials into the quantitative synthesis, three included more than one cluster randomization trials. The trials for each meta-analysis were mixed up with different randomized designs and units of analysis. These meta-analyses probably the ones that used sensible methods because the reviewers were aware of heterogeneity that might be due to cluster randomization trials.

Three meta-analyses that included cluster randomization trials with a mixture of different randomized designs and units of analysis, attempted to adjust for clustering effect in the quantitative synthesis. Details of adjustment for each meta-analysis were presented in Table 3.4. The outcome measures of these three meta-analyses were binary data. The meta-analyses had individual explanation for the clustering effect adjustment in the following three paragraphs.

First was the meta-analysis evaluating the value of mammographic screening for women under 50 years of age (45). It included six individually randomization trials and two cluster randomization trials. For the two cluster randomization trials, one (46) used the design of stratified randomization and individual level as the unit of analysis adjusted for clustering effect. The other (47) used matched-pair design and also individual level as the unit of analysis but ignored clustering effect. The applied technique of Mantel-Haenszel for clustered binary data, proposed by Rao and Scott (48), was used in the sensitivity analysis. The technique aimed to estimate an overall odds ratio of K 2 x 2 tables of independent clustered data in binary outcome. By using the Rao and Scott's method, each included trial of the meta-analysis was taken to represent an independent group of the clustered binary data. The method required clustering effect of each treatment group for each trial to be adjusted for in the analysis. Since there was less information on this process in the methodology part of the meta-analysis, it was unclear exactly how the authors managed this issue. But they reported that each of the two cluster randomization trial results was allowed for the same degree of clustering effect of relative 90 per cent (=100(1/design effect)) in the synthesis without any explanation on the adjustment. This might elevate the problem of inappropriate adjustment. It was because only one cluster randomization trial (46) reported the estimate of relative efficiency due to cluster sampling of 87 per cent. In addition, the six individually randomization trials seemed to be treated as having one cluster in each arm of the trial. This issue did not satisfy the requirement of the method that needed a large number of clusters in each arm of each trial to provide valid results. Thus the Rao and Scott's method would be

inappropriate for estimating an overall odds ratio of any meta-analysis including a mixture of individually and cluster randomization trials, which was the case of this meta-analysis.

Second was the meta-analysis assessing effect of vitamin A supplementation on child mortality(1). All eight trials included were cluster randomized. Six of them used a completely randomized design, one used matched-pair and the other reported unclear information on the randomized design. The analyses were reported at cluster level in one trial and at individual level in seven trials, of which three trials adjusted for clustering effect. The meta-analysis reported the common method of DerSimonian&Laird (49), which was the random effects model, used to estimate an overall odds ratio. Each pooled odds ratio was adjusted for clustering effect by increasing the variance with equal estimate of 30 per cent. The report presented that the figure was determined from some included cluster randomization trials that provided sufficient information on the clustering effect ranging from 10 to 44 per cent. In fact the cluster randomization trials were quite different in terms of types of unit of treatment allocation like wards, household, clusters, villages, districts areas and slums, and number of clusters of each trial. Thus it seemed to be unfair to account for clustering effect with the same degree for individual pooled odds ratio. In addition, some results of the cluster randomization trials (34, 50, 51) were already adjusted for clustering effect, and one (52) had the result at cluster level. The approach of adjustment for clustering effects used in this meta-analysis might be reasonable if the trials included have quite similar units of treatment allocation and number of clusters of each arm for each trial.

The third meta-analysis was on vitamin A supplementation on childhood pneumonia mortality(53). This meta-analysis included five cluster randomization trials, four(34, 50-52) of which overlapped with trials of former meta-analysis (1). Four of these five cluster randomization trials used a completely randomized design and one used a matched-pair design. Three of the five trials reported analyses performed at individual level, two of them adjusted for clustering effect. The remaining two trials reported analyses done at cluster level. The meta-analysis reported the fixed effect model of Mantel-Haenszel method used to pool the results. Individual pooled results were adjusted for clustering effect by increasing the variances of their odds ratios with different degrees. The estimates of the adjusted effects were obtained from the meta-analysis studied by Beaton et al. (54), which was done in a related topic to this meta-analysis. We did not review the Beaton et al.'s study (54) because it could not be accessed from any electronic database searched by our study. However, Donner et al.(55)mentioned that Beaton et al. (54) used the method of Rao and Scott (48) in their meta-analysis with satisfaction to the method assumption. The adjustment for different degrees of clustering effects seemed to be a reasonable procedure because the unit of randomization for each cluster randomization trial was quite different. But there were two trials(52, 53) that had the results analysed at cluster level and whether they needed to be adjusted for clustering effects. Therefore, if excluding the two trials, the adjustment approach shown in this meta-analysis seemed to be justified.

3.4 Comprehension of review results

In principle, when doing a meta-analysis including individually randomization trial results, an overall treatment effect could be estimated in a straightforward way, if the valid estimated treatment effects and their variances were provided. This concept could also be further applied to the meta-analyses that included results from cluster randomization trials with the same randomized designs and analyzed at individual level adjusted for clustering effect or at cluster level. Furthermore, even if the cluster randomization trials results were analyzed at individual level not adjusted for clustering effect, if all information on appropriate clustering effects was available, the results could be pooled. In practice this was unlikely to happen as seen in this review.

One simple approach for adjustment of clustering effect in binary outcome was the approach of Mantel-Haenszel proposed for clustered data by Rao and Scott (48). This approach could be applied to the meta-analysis of cluster randomization trials comparing two treatment groups with a completely randomized design(56). Requirements of the approach that relate to the results of cluster randomization trials were the results analysed at individual level. In addition, total sample size, count number of treatment outcome and clustering effect of each treatment group were needed. Furthermore, the method required a large number of clusters for each treatment group of each of individual cluster randomization trials. It might be impossible, however, to use this approach in real situations, because all the data required estimating an overall odds ratio by the approach were unlikely to be available.

The results show that 44.0 per cent (11/25) of the meta-analyses reported the methods considering the clustering effect in the synthesis. This figure was quite low. In addition, the meta-analyses that reported estimation approaches adjusted for clustering effect might provide imprecise estimates of overall treatment effects. Various issues needed to be considered.

It was found that 15 meta-analyses included more than one cluster randomization trials. The trials included in each of the meta-analyses had various randomizied designs as shown in Table 3. This was an additional source of heterogeneity and might raise more difficulties in methodologic issues beyond those already existed in meta-analysis including only individually randomization trials. The conventional approaches might be inappropriately used for estimating overall treatment effects from these trial results. However this issue was not considered properly in any meta-analysis reviewed and might lead to inappropriate use of the synthesis procedures. This difficulty would possibly produce imprecise result of the overall treatment effect.

Invalid results obtained from cluster randomization trials, which were the results without adjusting for clustering effect, were crucial and lead to a difficulty in estimating the effects in the meta-analysis including the trial results, especially when the trials did not report clustering effect information.

The figure of 56.8 per cent (42/74) of the cluster randomization trial results that adjusted for clustering effect was found in this review. It was interesting that the results reflected the persisting figure on analysis of cluster randomization trials, compared to the reviews by Donner A, et al. in 1990 on cluster randomized non-therapeutic intervention trials from 1979-1989(41), and later by Simpson JM, et al. in 1995 on cluster randomized primary prevention trials from 1990-1993 (32). They found that 50.0 per cent (8/16) and 57.1 per cent (12/21) respectively, took account of clustering effect in the analyses. One reason might be that the cluster randomization trials reviewed in this study were performed around the same period as those of the previous reviews. In addition, three cluster randomization trials(34, 47, 57) in the previous reviews were included in this study.

Recently, some authors(58-60) have proposed to report design effects and intra cluster correlation when publishing cluster randomization trials. Thus, hopefully, the difficulty situation as mentioned above would be corrected in the near future.

There were 52 per cent (13/25) of the meta-analyses used inappropriate methods that ignored clustering effect to combine invalid results of cluster randomization trials. Here, we could speculate about the reasons. First, as 9 out of 13 meta-analyses were obtained from the Cochrane library. The Cochrane collaboration lacked the appropriate software to analyse the cluster randomization trial results during the study period. Some authors were aware of this constraint and warned readers that the confidence intervals provided might be too narrow. Second, there was generally neither guideline nor proposal methods to combine cluster randomization trial results. Finally, some meta-analysts might not know that variation of the estimated outcome obtained from the cluster randomization trials differed from that of the

individually randomization trials and that would have an impact on the combined results. However, recently some approaches involving binary outcome variable have been proposed by Donner et al(55, 56).

The results show three meta-analyses(1, 45, 53) involving binary endpoints attempted to take clustering effect into account in the analysis. They were done to solve the problem of invalid results. The invalid results were due to not adjust for clustering effect in the analysis at individual level. The synthesis attempted to estimate the clustering effects; some from internal available clustering effect information and some from external clustering effects. Some unclear issues were still found. First, no rationale for the methods used to estimate clustering effects was seen. Second, some cluster randomization trials providing results with appropriate analysis seemed to be forced to adjust for clustering effect. Third, complex situations; different randomized designs, heterogeneity in units of randomization and variation of the randomization units, and different levels of units of analysis among the cluster randomization trials included were found but not taken into account in the three meta-analyses.

Some limitations of the review are considered. One meta-analysis (54) satisfied inclusion criteria was not reviewed because we could not retrieved it from the searched electronic database. It is, however, mentioned in Donner at al.(55) that the Rao and Scott's method was used in the meta-analysis. The method is not different from what we found in the review. In addition, four incorrect references of cluster randomization trials could not be accessed. With these limitations we believe the finding of this review could reflect the recent practice of meta-analyses involving cluster randomization.

From the difficulties found in the reviewed meta-analysis involving cluster randomization trials, some suggestions are introduced. The first suggestion focuses on some specific issues in reporting cluster randomization trials that relate to the information needed in meta-analysis. Number of clusters assigned to each treatment group is required in the report. This is because when the trial has only one cluster for each treatment arm, variation between clusters, even exists, is confounded by the treatment effect and cannot be measured from the trial(60). Consequently, including this trial in a meta-analysis, there is a need to adjust for clustering effect from similar available source. Unit of analysis must be clearly stated whether at cluster or individual level. If analysis is performed at individual level, the degree of clustering effects for each treatment group that is adjusted for in the analysis must be reported. This information is benefit not only to the meta-analysis where the trial is included, but also to any future plan for performing a cluster randomization trial in related field. There, however, have been more complete suggestion for reporting the trials provided by Donner and Klar (61), and Elbourne and Campbell (59).

The second suggestion focuses on the synthesis approach. If the number of cluster randomization trials included is relatively small and diverse in randomized designs and units, it might be reasonable to do qualitative synthesis, i.e. explaining individual cluster randomization trials separately as was done in some reviewed meta-analyses (42, 62-66). Alternatively, if number of the trials is large, subgroup analyses, which are meta-analyses on subgroups of the studies, might be sensible when the categories of interest factors are quite small, like three types of randomized designs: completely randomized, matched pair and stratified randomized. Some approaches involving binary outcome variable have been proposed by Donner et al.(55, 56). They are recommended to be used for the included trials involve a completely randomized design. Advantages and disadvantages of each approach are also provided. In addition, recommendations of application of the approaches to combine results from different designs under limitation issues have also been discussed in the literature (55).

In conclusion, attempts to work on some difficulties due to involving cluster randomization trials in meta-analysis were seen. Some suggestions on the methods for meta-analyses of cluster randomization trials measured in a binary outcome have been proposed (55, 56). The problem of heterogeneity results, from complex situations on various randomized designs and units, different eligibility criteria at cluster and individual level, and unit of analysis that might be beyond the heterogeneity results obtained from individually randomization trials, have been found and still needed further methodologic investigation.

Table 3.2 Numbers of individually and cluster randomization trials included for individual meta-analyses reviewed

Meta-analysis reference	Type of Intervention	Number of trials included	
		individually randomization	cluster randomization
(67)	Health Care	1	(1)*
(62)	Health Care	1	1[1]
(68)	Health Care	6	1[2]
(69)	Health Care	6	1[3]
(43)	Screening	7	1[4]
(44)	Health Care	13	1[5]
(63)	Health Care	14	1[6]
(70)	Health Care	16	1[7]
(71)	Health Care	28	1[8]
(64)	Educational	39	1[9]
(45)	Screening	6	2[4-10]
(42)	Educational	2	3[11-13]
(72)	Educational	13	3[14-16]
(65)	Educational	34	3[17-19]
(73)	Educational	38	3[13,20,21]
(74)	Health Care	10	4[22-25]
(75)	Educational	15	4[20,26-28]
(76)	Health Care	23	4[29-32]
(77)	Health Care	-	5[33-37]
(53)	Health Care	-	5[35-39]
(78)	Health Care	13	5[40-44]
(66)	Educational	2	6[17,32,45-48]
(79)	Health Care	3	7[49-54](1)**
(1)	Health Care	-	8[33-39,55]
(80)	Health Care	1	17[49-52,56-62](3)**
Total		297	89 (5)

Numbers in the parentheses were papers on cluster randomization trials, for which original papers cannot be retrieved.
Superscript is number of references, presented in Appendix
* Paper is in Russian, its detail was extracted from the meta-analysis
** Missing papers as incorrectly referenced

Table 3.3 Details of individual cluster randomized trials for each meta-analysis in terms of randomization design and analysis level, and combining method of meta-analysis

Meta-analysis reference	No. CRT@ included	Randomization design	Analysis level	Combining method
(36)	1	1 C	1 IU	T
(33)	1	1 C	1 IU	T
(5)	1	1 S	1 IA	T*
(32)	1	1 U	1 U	T
(4)	1	1 S	1 IA	T*
(35)	1	1 S	1 C	T
(37)	3	1 C, 1M, 1 U	2 IU, 1 U	T
(38)	3	1 S, 1 U	1 IA, 2 U	T
(39)	4	1 C, 3 M	2 C, 2 IU	T
(40)	4	1 C, 1 M, 2 U	1 C, 1 IA, 1 IU, 1 U	T
(41)	4	2 C, 1 S, 1 U	1 IA, 3 IU	T
(16)	5	4 C, 1 M	1 C, 3 IA, 1 IU	T
(43)	5	3 C, 1 M, 1 S	2 IA, 3 IU	T
(44)	6	1 C, 2 M, 3 S	1 C, 5 IU	T
(45)	14	4 C, 6 M, 4 S	9 C, 5 IU	T
(26)	1	1 M	1 C	D*
(27)	1	1 C	1 C	D*
(28)	1	1 S	1 IA	D*
(3)	3	2 C, 1 M	3 IA	D*
(29)	3	3 C	2 C, 1 IA	D*
(30)	6	5 C, 1 M	1 C, 2 IA, 3 U	D*
(6)	2	1 M, 1 S	1 IA, 1 IU	A
(42)	5	4 C, 1 M	2 C, 2 IA, 1 IU	A
(10)	8	6 C, 1 M, 1 U	1 C, 3 IA, 4 IU	A
(34)	1	1 M	1 C	U
Total	85			

Randomization designs
- C completely randomized
- M matched-pair
- S stratified randomized

Analysis level
- C cluster
- IA individual adjusted for clustering effect
- IU individual unadjusted for clustering effect
- U unclear

Combining method
- A account for clustering effect
- D describe CRT results separately
- T treated CRT results as if of IRT and use fixed effect models
- U unclear method

- * Reasonable method
- @ Cluster randomized trial

Table3.4 Design of randomization and analyses level of the individual included CRT@ of the 3 meta-analyses managing clustering effect in the combination

A. Meta-analysis on mamographic screening trials (45)

Number of CRTs@ reviewed				Management of clustering effect in the combination
Randomized design		Analyses level		
Stratified	1	Adjusted at individual level	1	Proposed method of Mantel-Haenszel by Rao & Scott [9] for clustered binary data is used in a sensitivity analysis to examine the clustering effect of the two including CRTs@
Matched-pair	1	Unadjusted at individual level	1	
Total	2	Total	2	

B. Meta-analysis on vitamin A supplementation (1)

Number of CRTs@ reviewed				Management of clustering effect in the combination
Randomized design		Analysis level		
Completely randomized	6	Adjusted at individual level	3	Dersimonian&Laird method[11] adjusted for clustering effect by increasing variance of each pooled log-odds ratio with a fixed estimate of 30 %. The estimate is determined from some included CRTs@ which provided sufficient clustering effect.
		Unadjusted at individual level	4	
Matched-pair	1			
		Cluster level	1	
Unclear	1			
Total	8	Total	8	

C. Meta-analysis on vitamin A supplementation (53)

Number of CRTs@ reviewed				Management of clustering effect in the combination
Randomized design		Analysis level		
Completely randomized	4	Adjusted at individual level	2	Mantel-Haenszel method adjusted for clustering effect for each pooled result differently. The adjusted effects are estimated from the external CRT@ study done in a similar topic to the including CRTs@
		Unadjusted at individual level	1	
Matched-pair	1			
		Cluster level	2	
Total	5	Total	5	

@ Cluster randomization trial

Topic 4: Examples of published meta analyses involving cluster randomized trials in binary outcome

Three published meta analyses involving cluster randomized trials in different situations are used as examples in this study. The first is a meta analysis to evaluate the effect of vitamin A supplementation on child mortality (1). All eight trials included are cluster randomized trials. The second meta analysis investigates the effect of mammographic screening on breast cancer mortality in women under 50 years(2, 81). It includes eight trials. Five of the trials included are cluster randomized trials. The third meta analysis examines effects of multiple risk factor interventions on primary intervention of coronary heart disease(3). This meta-analysis includes fourteen trials. Five of the trials randomly allocate treatments to groups or clusters of subjects.

4.1 Meta-analysis of vitamin A supplementation trials

This meta-analysis(1) includes eight community-based trials. They were performed in developing countries to examine the relationship of vitamin A supplementation and mortality in children aged 6 to 72 months, which is in a wide age range. None of the trials assigned individual children to treatment groups. The units of treatment allocation vary between trials The treatments are vitamin A supplementation and control group. The control groups also vary from trial to trial. Detail of treatments for individual trials are presented in the original meta-analysis(1). Here, some design characteristics and summary statistics of the outcome are shown in table 4.1. Five of the eight trials provide mean follow- up periods of the children for 12 months. One trial observed the children for a mean period of 42 months. Each trial has similar sample sizes for each treatment group. The trial sizes of children studied ranged from 3,428 to 28,740 with a mean of 12,399.

Child mortality is the outcome in each trial. Details like number of child deaths and children assigned to each treatment group of individual trials are also presented in table 4.1. The observed odds ratio range from 0.20 to 1.04. The meta-analysis reports DerSimonian and Laird pooled overall odds ratio of 0.70 (95 per cent CI 0.58 to 0.85).

Data available for reanalysing this meta-analysis are the number of child deaths, total number of children assigned and the mean follow-up period in months for each trial. Since the trials have different mean follow-up periods, reanalysis in the next four chapters is performed using relative risk adjusted for mean follow-up period. Details of the calculation is described in topic 5. The same procedure is done for the next two examples.

4.2 Meta-analysis of mammographic screening trials

The meta-analysis(2) is performed to evaluate the effect of mammographic screening on reduction of breast cancer mortality in women aged less than 50 years.

The meta analysis includes 8 identified trials perfomed on women from various western countries. The treatments are mammographic screening and a control group. Four of the trials included randomly assigned treatments to clusters of women. The clusters consist of various types such as area, practice, birth date and birth year. The remaining four trials had randomly assigned treatments to individual women. One trial presents 18 per cent of the subjects as cluster randomized by birth date, and the other 82 per cent as individually assigned to receive treatments. Because individual randomization occurred in most of the women studied, the trial is classified here as an individully randomized trial. The trials are also different in other design characteristics, like the population studied, contamination rates (unlikely up to 51%), and mean follow-up periods. The mean follow-up periods range from 10 to 18 years. This information is extracted from another paper (81) that included the same

eight trials. In addition, radiation dose per breast and blinding process are different among trials. Some design characteristics are presented in table 4.2.

The primary outcome is breast cancer mortality. Details of number of breast cancer deaths and total number of women in each treatment group are also shown in the table. Two trials, trial 4 and 6, show high imbalance of women studied for each treatment group. The trial sizes range from 25,941 to 89,835 with a mean of 57,044. The observed relative risks range from 0.55 to 1.08.

The meta-analysis reports two fixed effect pooled relative risk estimates. The first is 1.04 (95 per cent CI 0.84 to 1.27) for the two trials, trial 2 and 3, with adequate randomization methods and baseline comparability. The second pooled relative risk is 0.75 (95 per cent CI 0.67 to 0.83) for the other six trials that had not been adequately randomized and had more favorable outcomes for screening than those two trials. Their results were homogeneous ($p=0.23$ for test of heterogeneity). This estimate is significantly different from that for the two adequately randomized trials ($z=2.60$, $p=0.005$).

Data available for reanalysing this meta-analysis are the estimated relative risk that are recalculated with adjustment for mean follow-up periods in years, and treatment allocation, called randomization design in this thesis.

4.3 Meta-analysis of multiple risk factor interventions trials

The meta analysis (3) was done to assess the effectiveness of multiple risk factor interventions to reduce cardiovascular risk factors from coronary heart disease. Study subjects are adults aged at least 40 years and having no clinical evidence of established cardiovascular disease.

A total of 18 trials are included, of which thirteen randomly assigned interventions to individual subjects. The other five trials assigned interventions to different groups of individuals, such as families, households, worksites and factories. The interventions included counselling or educational approaches with or without pharmacological interventions aimed to reduce more than one cardiovascular risk factor.

Outcome measures were total mortality, coronary heart disease(CHD) mortality, net change in blood pressure, smoking and total blood cholesterol. Some measures were not available in some trials. Hence, various number of trials could be analysed for individual outcome measures. In this thesis reanalysis is performed only for smoking prevalence as it is the outcome that could be analysed from the biggest number of trials, which is 14 trials. Details of some trial designs and the results of individual trials are presented in table 4.3.

Duration of the interventions vary between trials and range from 1 to 11.5 years. One trial, trial 4, shows a very remarkable imbalanced number of subjects between intervention and control group; 16,908 for intervention and 1,902 for the control. The trial sizes range from 335 to 18,810 with a mean of 3674. The observed odds ratio for individual trials range from 0.57 to 1.12. The meta-analysis reports an overall odds reduction of 16 per cent (95 per cent CI 3 to 27 per cents) using a random effects model.

The reanalysis of this meta-analysis is based on the recalculated relative risk adjusted for duration of the interventions, to adjust for varing.

Table 4.1* Design characteristics and summary statistics of outcome for individual trials

Trial	Location Published years	Treatment unit, No.	Follow up Period, (months)	No.deaths Total		Odds Ratio (95per cent CI)
				Vitamin A	Control	
1	Aceh,Indonesia 1986	Village, 450	12	101 12,991	130 12,209	0.73 (0.56 to 0.95)
2	Java, Indonesia 1988	Area, 2	12	186 5,775	250 5,445	0.69 (0.57 to 0.84)
3	Hyderabad, India 1990	Village, 84	12	39 7,691	41 8,084	1.0 (0.64 to 1.55)
4	Tamil Nadu, India 1990	Cluster, 206	12	37 7,764	80 7,655	0.45 (0.31 to 0.67)
5	Bombay, India 1991	Slum, 2	42	7 1,784	32 1,644	0.20 (0.09 to 0.45)
6	Jumla, Nepal 1991	District, 16	5	138 3,786	167 3,411	0.73 (0.58 to 0.93)
7	Sarlahi, Nepal, 1991	Ward, 261	12	152 14,487	210 14,143	0.70 (0.57 to 0.87)
8	Northern, Sudan 1992	Household, 16,789	18	123 14,446	117 14,294	1.04 (0.81 to 1.34)

* data is modified from table 4 in Fawzi, 1993(1)

Table 4.2* Design characteristics and summary statistics of outcome for individual trials

Trial	Location Published years	Treatment allocation	Follow up Period, (years)	No.death Total		Relative risk (95 per cent CI)
				Screen	Control	
1	New York 1988	Age-matched random	18.0	153 30,131	196 30,565	0.79 (0.64 to 0.98)
2	Malmo 1988	Cluster by birth date	10.0	63 21,088	66 21,195	0.96 (0.68 to 1.35)
3	Canadian 1997	Individual	10.5	120 44,925	111 44,910	1.08 (0.84 to 1.40)
4	Kopparberg 1995	Cluster by area	13.0	126 38,589	104 18,582	0.58 (0.45 to 0.76)
5	Ostergotland 1995	Cluster by area	13.0	135 38,491	173 37,403	0.76 (0.61 to0.95)
6	Stockholm 1997	Cluster by birth date	11.4	66 40,318	45 19,943	0.73 (0.50 to 1.06)
7	Goteborg 1997	18%cluster by birth date 82% individual	10.0	18 11,724	40 14,217	0.55 (0.31 to 0.95)
8	Edinburgh 1999	Cluster by practice	14.0	156 22,926	167 21,342	0.87 (0.70 to 1.08)

* data is modified from table 2 of Gotzsche,et al in 2000(2) and from table 2 of Ringash in 2001(81)

Table 4.3* Design characteristics and summary statistics of outcome for individual trials

No.	Trial, Published years	Treatment unit	Interven-tion period (years)	Smoking prevalence Total		Odds Ratio (95 per cent CI)
				Intervention	Control	
1	MRFIT Study, 1982	Individual men	6.0	1,847 5,754	2,554 5,638	0.57 (0.53 to 0.62)
2	Gothenberg Study,1986	Individual men	11.8	691 1,473	699 1,404	0.89 (0.77 to 1.03)
3	Oslo Study, 1986	Individuals	5.0	428 604	496 628	0.65 (0.50 to 0.84)
4	WHO Factories,1989	Factories	6.0	7,910 16,908	897 1,902	0.98 (0.90 to 1.08)
5	Abingdon, 1990	Individual adults	1.0	46 168	42 167	1.12 (0.69 to1.82)
6	Tromso men, 1991	Individual men	6.0	247 525	284 535	0.79 (0.62 to 1.00)
7	Tromso wives, 1991	Individual women	6.0	186 422	178 387	0.93 (0.70 to 1.22)
8	Family Heart -men, 1994	Families	1.0	337 1,767	500 2,174	0.79 (0.68 to 0.92)
9	Family Heart -women,1994	Households	1.0	215 1,217	301 1,402	0.79 (0.65 to 0.95)
10	OXCHECK Study,1994	Households	3.0	544 2,205	506 1,916	0.91 (0.79 to 1.05)
11	Swedish RIS, 1994	Individual men	3.0	55 253	70 255	0.74 (0.49 to 1.10)
12	CELL Study, 1995	Individual adults	1.0	139 292	148 310	0.99 (0.72 to 1.37)
13	Finnish men, 1995	Volunteer men	5.0	125 575	131 580	0.95 (0.72 to 1.26)
14	Take heart, 1995	Worksites	1.5	190 1,057	166 920	1.00 (0.79 to 1.25)

* data is modified from the table of outcome: Smoking prevalence of Ebrahim et al. in 2001 (3)

Topic 5: Meta-analysis: some essential issues in quantitative synthesis and simple conventional approaches

5.1 Introduction

Meta-analysis is well accepted as an essential tool used to evaluate health care. One of the main aims of meta-analysis is to produce a more accurate estimate of the treatment effect of interest than the estimate obtained by possibly using a single trial. As different trials are conducted using different populations, different designs and a whole range of other trial-specific factors, it has been suggested that pooling them would yield an estimate that has wider generalizability than any single trial. In addition, by performing a meta-analysis, it may be possible to explain the differences between results from individual trials (82).

Once meta-analysts have finished with their critical appraisal of the primary trials, they will be involved in the statistical analysis. This topic provides some essential elements of the analysis and also describes the estimation of simple conventional approaches. Section 5.2 discusses the quantification of treatment effects required from individual trials and scaling of the treatment effects. Section 5.3 presents a discussion on heterogeneity. Section 5.4 discusses publication bias. Section 5.5 discusses modelling of variation and simple conventional approaches of the models. It also discusses some extension methods used to overcome the problem of a constant variance to measure heterogeneity between trials.

5.2 Quantitative requirement and scaling of treatment effects

To obtain a numerical conclusion from a meta-analysis, the same measurement scale of observed treatment effects with their variances from individual trials is needed. However, it is hardly possible to obtain such information from published trials(15). The trials may present their treatment effect in different ways. An example is found in the meta analysis of 6 studies performed to investigate the effect of screening for diseases on the levels of psychological morbidity (Chapter2;pp29-63 in(17)). The level of long-term anxiety is considered as the outcome measure of the screening programmes. The difference of mean change between treatment and control groups is the observed treatment effect. However, the meta analysis presents only 2 studies measuring the treatment effects on this scale. The other 4 studies present means of each treatment group at baseline and after a follow-up period. In addition, some information, such as the variance, may not be available. Other information available, like p value, may be used to obtain the missing information. However, the original authors of the trials are usually requested to provide the necessary information(15).

In medical research, effect of treatment is often measured in binary outcomes, such as dead/alive and response/non-response. Risk difference (often called absolute risk reduction), relative risk (risk ratio) and odds ratio are common measures of such effect. The risk difference is more common in clinical trials. This measure is easily understood and interpreted. It is also usually used to estimate sample size. Nevertheless, the risk difference is not a common one in meta-analysis. A discussion on this issue is presented in the following paragraph.

It is still a debatable issue as to which measures should be used in meta-analyses(10). Even any measure can be used, the absolute risk is likely to yield severe heterogeneity of treatment effects across trials(10, 15). This is because the absolute risk is based on the underlying risk of the study subjects of individual trials(11, 15). Relative measure is more common in meta-analysis. However, between the two relative measures of odds and risk ratios, the odds ratio is more common than the risk ratio or relative risk. This is because the odds ratio is the measure of most statistical methods available in meta-analyses(15). Eventually, some literature showing that both odds ratio and relative risk are equally likely to present less heterogeneity of treatment effects across trials(10).

The treatment effect is frequently measured as a continuous outcome in medical research. Examples are the change of blood pressure in hypertension patients and body mass index in malnourished children. Mean difference is a common measure for continuous outcome. However, there may be different trials that measure effect of the same treament but in a different scale. For example, some trials may measure birth weight in units of pound while others may measure it in units of gram. Standardized difference, sometimes called effect size, is always suggested for comparing such trials in meta-analysis. The standardized difference is calculated from a ratio of means difference to its standard deviation.

This study involves examples of three published meta-analyses of trials that measure treatment effect in binary outcome with different follow-up periods as presented in Topic 4. Although the relative risk is used as a measure of treatment effect, the natural logarithm tranformation of relative risk is used for the purpose of combining via meta-analysis(11). A formal description of the calculations of a relative risk, log-relative risk and variance of log-relative risk from data of individual trials are presented in table 5.1.

Table 5.1 Summary statistics of I individual trials for two treatments comparison in binary outcome with different mean follow-up periods in a meta-analysis

Trial	Mean follow-up period *	Groups	Frequencies of event #	Total	Event rate	Relative risk (RR)	Log-RR ($\hat{\theta}$)	Var[Log-RR] Var($\hat{\theta}$)
1	T_1	Treatment Control	A_1 B_1	M_1 N_1	$P_1 = A_1/M_1T_1$ $Q_1 = B_1/N_1T_1$	P_1/Q_1	$\text{Log}(P_1/Q_1)$	$\frac{1}{A_1}+\frac{1}{B_1}$
2	T_2	Treatment Control	A_2 B_2	M_2 N_2	$P_2 = A_2/M_2T_2$ $Q_2 = B_2/N_2T_2$	P_2/Q_2	$\text{Log}(P_2/Q_2)$	$\frac{1}{A_2}+\frac{1}{B_2}$
.	.		.	.	.	.	.	.
.	.		.	.	.	.	.	.
I	T_I	Treatment Control	A_I B_I	M_I N_I	$P_I = A_I/M_IT_I$ $Q_I = B_I/N_IT_I$	P_I/Q_I	$\text{Log}(P_I/Q_I)$	$\frac{1}{A_I}+\frac{1}{B_I}$

* in any units of time, e.g. years, months, etc. # e.g. number of children deaths, etc.

5.3 Heterogeneity issue and testing

Different studies are likely to yield different treatment effects even when they address similar scientific questions. In meta-analysis, difference of treatment effects across trials beyond chance alone (sampling error), usually called *'heterogeneity'*, is a very important issue. Information of size of heterogeneity not only leads to a choice of pooling methods but also affects the conclusion of meta-analysis.

The heterogeneity of treatment effects across trials may be due to some biases in the difference in trial designs, treatment procedures and subject characteristics between trials, and unobserved random effects (83, 84). When difference of treatment effects across trials is due to chance alone, sampling error or within trial variation is solely the variation allowed in the synthesis method. This is well known as the fixed effect model described in section 5.5. When heterogeneity is detected, identification of sources of the heterogeneity is needed. Potential characteristics of subject and trial levels are one possible source of systematic error that must be considered. Another possibility is to consider the random effects of between trials. This potential variability needs to be taken into account in the synthesis methods.

The challenge is how to investigate and quantify heterogeneity. A test for heterogeneity in form of the Q statistic is a common statistical tool used to investigate the evidence of heterogeneity results across trials. The Q statistic is valid for both binary and continuous outcome. To perform the test, a null hypothesis of homogeneity results across trials is set, which is

$H_o : \theta_i = \theta$ for all i where θ_i is the true treatment effect of trial i and θ is a common true treatment effect,

or $H_o : \tau^2 = 0$ where τ^2 is the variance of random effects between trials.

It versus an alternative hypothesis that there is at least one θ_i different, or $H_A : \tau^2 > 0$.

Under H_o, for a large sample sizes,

$$Q = \sum_{i=1}^{I} w_i[\hat{\theta}_i - \hat{\theta}]^2 \qquad (5.1)$$

It has an approximate χ^2_{I-1} distribution where $\hat{\theta}$ is a weighted average of estimated overall treatment effect, $\hat{\theta} = \sum_{i=1}^{I} w_i \hat{\theta}_i \Big/ \sum_{i=1}^{I} w_i$, and w_i is the weight given to trial i. Here, w_i is the inverse variance of the observed treatment effect, i.e. $w_i = [1/\hat{\sigma}_i^2]$.

The H_o would be rejected when the Q statistic exceeds the upper-tail critical value of χ^2_{I-1} distribution. It means that the variance of the treatment effects between trials is significantly greater than what would be expected by sampling error when all trials estimate the same underlying treatment effect. Hence, we can conclude that there is evidence of heterogeneity(13, 85). This conclusion leads to the choice of random effects model in the synthesis. Alternatively, the choice of fixed effects model is introduced in the synthesis if no further evidence to support heterogeneity is found. Application of this test is illustrated in section 5.6.

This test is, however, appropriate when a fixed number of trials with large trial sizes are available. Some limitations of the Q statistics have been pointed out (6, 7, 15, 84, 86-89). A main concern is that power of the test can be low, especially in the case of sparse data or when one trial has much more precise estimate effect than the rest (90).

Fleiss(11) suggests using an arbitrary significant level at 10 per cent rather than 5 per cent in the test for heterogeneity by Q statistic, because it would allow higher probability to detect significant heterogeneity. Some authors (85, 90) recommend the use of random effects model routinely to evaluate sensitivity of the fixed effect model. If the results of overall treatment effect from the two methods are likely to be similar, and the variance of random effects between trials is close to zero ($\tau^2 \cong 0$), the conclusion of fixed effect approaches would be accepted.

Hardy and Thompson (89) suggest using likelihood ratio statistic based on the marginal likelihood of each trial to test for the null hypothesis of homogeneous treatment effects across trials ($H_o : \tau^2 = 0$). The likelihood ratio statistic (LRT) is expressed as(89)

$$-2\{LL(0) - LL(\hat{\tau}^2)\}$$

where $LL(0)$ is the log-likelihood for $\tau^2 = 0$ under null hypothesis of homogeneity, and

$LL(\hat{\tau}^2)$ is the maximum log-likelihood for $\tau^2 \geq 0$.

Since $H_0 : \tau^2 = 0$ is part of the boundary of $H_A : \tau^2 > 0$, the asymptotic distribution of the likelihood ratio test under H_0 is no longer chi-square at 1 degree of freedom. In this case, it is shown (91) that the square root of LRT can be compared to a one-tailed standard normal distribution. This implied that the p value for the likelihood ratio test can be obtained from ½ of the probability of LRT on the chi-square distribution with 1 degree of freedom.

Application of this test will be presented in the next topic. Also, the results obtained from the regression models in relation to the results obtained from the Q statistic will be discussed. Some other tests have been suggested but they are all appropriate only to the odds ratio (90).

5.4 Publication bias

Meta-analysis relies heavily on the published literature. However, the studies included in the literature may not represent all relevant studies conducted on the topic of interest. There is empirical evidence (92) that statistical significance is an important key issue of publication. Studies with non-significant results may not be submitted by investigators for publication, and editors may not publish studies with non-significant results even if they were submitted (93). Therefore, studies with statistical significant or interesting results are potentially more likely to appear in the published literature. This is the cause of publication bias.

The evidence of publication bias is well demonstrated. When meta-analysis includes only the published literature, this can potentially lead to biased over-optimistic conclusions (94). These biased conclusions may have a further impact on health policy, clinical decision and outcome of patient management. However, there is little empirical evidence on the latter issue(94).

Attempts have been made to alleviate the problem of publication bias, such as by encouraging publication of previously unpublished trials(95), and the establishment of registries for the prospective registration trials (96). This issue is currently a big concern among meta-analysts (82). Searching for relevant unpublished trials is the other attempt to improve publication bias. However, it may be difficult to identify such trials. Therefore, many authors (85, 97-99) encourage routine assessment publication bias based on available data in the meta-analysis.

The funnel plot, proposed by Light and Pillemer in 1984 (100), is perhaps the most common method used to informally identify the existence of publication bias in meta-analysis(85, 101). To construct a funnel plot, each trial-specific effect is plotted against a measure of its precision. Precision may be defined differently according to the inverse of standard error of the estimate effect or the trial sizes. An interpretation of the funnel plot is given(100, 102) that if there is no publication bias, the plots are symmetrically distributed around the overall effect in the shape of an inverted funnel. Alternatively, the plots will become asymmetrical and the overall effect of the meta-analysis will be biased. However, there are some disputes concerning the asymmetry of the funnel plot. It is argued that this may be due to other factors such as real heterogeneity(103) and heterogeneity in treatment effect between low and high risk groups(104). Furthermore Tang, et al(103) show that shapes of the funnel plots are different according to different definition of precision. So, because of this evidence, care should be taken when interpreting asymmetrical funnel plots.

Some comparative methods of testing for bias have been proposed to detect publication bias. They are based on symmetry assumptions as a funnel plot assessment. The first method is the rank correlation method using Kendall's tau to evaluate the association between the effect and variance of the treatment effect(105). The second is the simple linear regression of the standardized estimate of treatment effect on the precision of the estimate (102). The third is the regression of the treatment effect on sample size(99).

There is no standard method on further process after assessing publication bias. However, some methods to investigate impact of publication bias on the estimate of overall treatment effect have been developed and proposed (95, 101, 106). Most of the methods, however, are still complex and difficult to implement(82). They are, therefore, not widely used. Nevertheless, these methods are beneficial tools to be used for sensitivity analysis.

5.5 Modelling of variation and simple conventional approaches

It is generally believed that treatment effects of different trials will be considered to differentiate with some level(85). The difference of treatment effects between trials is introduced in the synthesis process as quantity of variation. The variation is considered as fixed effect and random effects models.

The fixed effect model is assumed that each trial estimates the same underlying true treatment effect (9, 11, 85, 107). Here there is no consideration in variation between trials. Sampling error alone is considered to be the variation of treatment effect. Conversely, the random effects model assumes each trial to have its own underlying true treatment effect. But there is a distribution of all these underlying effects around a central value. Thus the overall variation is beyond the sampling error as it further accounts for the variation between trials. These two models lead to different pooling approaches.

To understand the concept of fixed effect and random effects models easily, this issue is discussed in conjunction with the simple conventional approaches. For the simple conventional approaches, *inverse variance weighted average* is the common method used for a large number of trials. The inverse variance weighted average produces an estimate of overall treatment effect from a weighted average of observed treatment effects across trials when each weight is given to each trial. The weight is usually obtained from an inverse variance of that trial. The approaches under fixed effect and random effects models are described as follows.

5.5.1 Fixed effect model

Assuming that there is a collection of I trials, i = 1, ..., I, each trial with a treatment group and a control group. Further let,

θ_i = a true treatment effect for trial i,

$\hat{\theta}_i$ = an observed treatment effect for trial i,

The $\hat{\theta}_i$, for example, can be the observed log-odds ratio or log-relative risk in a trial with binary outcome or the observed means difference in a trial with continuous outcome.

Under the fixed effect model, it is assumed that each observed treatment effect $\hat{\theta}_i$ is an estimate of unknown parameter θ_i, when $\theta_i=\theta$ for all i. This means all individual trials estimate the same true treatment effect. This is often called an 'assumption of homogeneity'. For individual large trials, a model is then specified by

$$\hat{\theta}_i = \theta + \varepsilon_i \qquad (5.2)$$

The ε_i represents a random deviation of each observed treatment effect $\hat{\theta}_i$ from the true treatment effect θ, and is assumed independent with mean zero and variance σ_i^2.

The observed treatment effects $\hat{\theta}_i$s, then, are asymptotically normally distributed and approximately unbiased (13), with a mean θ and variance σ_i^2.

In practice σ_i^2 is not known but it is generally estimated from each specific trial i. The σ_i^2 is often called within trial variance. In consequence, the estimated variance $\hat{\sigma}_i^2$ is then used to estimate both the true overall treatment effect and its associated variance.

An estimated weighted average of overall treatment effect is then estimated by means of the least squares method and expressed as

$$\hat{\theta} = \sum_{i=1}^{I} \hat{w}_i \hat{\theta}_i \Big/ \sum_{i=1}^{I} \hat{w}_i \qquad (5.3)$$

where $\hat{w}_i$ is the weight given to trial i. and $\hat{w}_i = [1/\hat{\sigma}_i^2]$.

Consequently, $\hat{\theta}$ is further has a mean θ and variance, $\text{var}(\hat{\theta}) = 1/\sum_{i=1}^{I} \hat{w}_i$, assuming an asymptotic normal distribution for $\hat{\theta}$. It allows calculating $100(1-\alpha)$ per cent confidence interval for θ. It is in the range of

$$\hat{\theta} \pm z_{(1-\frac{\alpha}{2})} (1/\sum_{i=1}^{I} \hat{w}_i)^{1/2} \qquad (5.4)$$

where $z_{(1-\frac{\alpha}{2})}$ is the $100(1-\frac{\alpha}{2})^{th}$ percentile value of the normal distribution.

The weighted average log-relative risk and the confidence interval are usually transformed back to be the relative risk by taking exponential log-relative risk.

Some other methods have been proposed to combine odds ratios in meta-analysis. They are the Mantel-Haenszel method, the Peto method and the Maximum likelihood estimator. Detail of the methods can be seen in many literature (11, 85, 108). These methods, however, provide very similar results for relatively large sample sizes(15, 109). Although the Peto method is a common one in the fixed effect model, it could produce serious underestimates or overestimates of the odds ratio in some extreme situations (11, 15, 108), e.g. in the trials with very unbalanced number of subjects between treatment groups.

As discussed in several literature(6, 7, 15, 17, 87-90), the fixed effect methods, which have a strong assumption of homogeneous treatment effect between trials, may not be appropriate in a real clinical practice. This is because it is mostly impossible to obtain an identical treatment effect from different trials, even when the same protocol is performed under different settings by different trained clinicians.

5.5.2 Random effects model

A random effects model is the one where treatment effects between different trials are different. This situation is usually named when there are heterogeneous treatment effects across trials. A random effects model supposes that each observed treatment effect $\hat{\theta}_i$ has its own distribution with a trial-specific mean θ_i and variance σ_i^2. Moreover, each θ_i is assumed to obtain from some super-population of treatment effects with mean θ and variance τ^2. This provides the two levels of hierarchical model

$$\hat{\theta}_i = \theta_i + \varepsilon_i \quad \text{where} \quad \text{var}(\varepsilon_i | \theta_i) = \sigma_i^2$$

and

$$\theta_i = \theta + u_i \quad \text{where} \quad \text{var}(u_i) = \tau^2 \qquad (5.5)$$

The u_i indicates random deviation of each-specific mean θ_i from the overall mean θ, and is assumed *to be independent from* ε_i. Thus θ and τ^2 represent the overall treatment

effect and between-trial variation, respectively. The variance τ^2 is usually known as a measure of the heterogeneity between trials(87).

The random effects model allows for both within trial variation σ_i^2 and between trial variation τ^2 to estimate the overall treatment effect and its variance. Thus, marginally the observed treatment effects are assumed normal with mean θ and variance $\sigma_i^2+\tau^2$.

Note that when $\tau^2 = 0$, the random effects model corresponds to the fixed effect model.

The weighted average treatment effect is then calculated in the same manner as in the fixed model. Here, the weight is allowed for an extra variation of between trials for each weight. The weight w_i^* that minimise the variance of $\hat{\theta}$ is an inverse variance of each trial, i.e.

$$w_i^* = \frac{1}{[\sigma_i^2 + \tau^2]}$$

In practice τ^2 is not known. The common approach to estimate τ^2 is the moment estimator given by DerSimonian and Laird(49). It is a non-iterative procedure and commonly used in meta-analysis. The estimated variance of between trials $\hat{\tau}^2$ is then given by equating Q statistic from (5.1) with its corresponding expected value, i.e.

$$\hat{\tau}^2 = \max[0, \frac{Q-(I-1)}{\sum_{i=1}^{I}\hat{w}_i - (\sum_{i=1}^{I}\hat{w}_i^2 \Big/ \sum_{i=1}^{I}\hat{w}_i)}] \qquad (5.6)$$

where $\hat{w}_i$ is the estimated weight provided by the fixed effect model. The estimated variance of between trials $\hat{\tau}^2$, then, gives the estimated weights $\hat{w}_i^* = 1/[(\hat{\sigma}_i^2 + \hat{\tau}^2]$. Thus, an estimate of weighted average treatment effect under the random effects model is expressed as

$$\hat{\theta}_{DL} = \sum_{i=1}^{I}\hat{w}_i^*\hat{\theta}_i \Big/ \sum_{i=1}^{I}\hat{w}_i^* \qquad (5.7)$$

The $\hat{\theta}_{DL}$ is assumed approximate normal, with a mean θ and variance $\frac{1}{\sum_{i=1}^{I}\hat{w}_i^*}$.

Then, an approximate $100(1-\alpha)$ per cent confidence interval for the true treatment effect θ is within a range

$$\hat{\theta}_{DL} \pm z_{(1-\frac{\alpha}{2})}(1/\sum_{i=1}^{I}\hat{w}_i^*)^{1/2} \qquad (5.8)$$

where $z_{(1-\frac{\alpha}{2})}$ is the $100(1-\frac{\alpha}{2})^{th}$ percentile value of the normal distribution.

The random effects approach allows for additional variation between trial τ^2 to estimate overall treatment effect and gives a wider confidence interval for the overall treatment effect as compared to the fixed effect approaches. This can be explained in the following. As the weights given to individual trial in the random effect methods, $\hat{w}_i^* = 1/[\hat{\sigma}_i^2 + \hat{\tau}^2]$ are generally lower than the weights of the fixed effect methods, $\hat{w}_i = [1/\hat{\sigma}_i^2]$. Consequently, the variance of estimated overall treatment effect $1/\sum_{i=1}^{I}\hat{w}_i^*$ in

the random effects approach is definitely greater than that, $1/\sum_{i=1}^{I}\hat{w}_i$, in the fixed effect approach.

However, there are some concerns over the approach of random effects model(87). The first concern is that the validity of normality assumption of the random effects remains questionable. The second is the difficulty in verifying normality assumption of random effects for meta-analyses. Third, is the inclusion of random effects to the estimation, only the single value of estimated variance between trials $\hat{\tau}^2$ is added into the weight. The model does not take into account the uncertainty associated with the estimate of the variance τ^2 . Thus, the given confidence intervals for the overall treatment effect may remain insufficiently conservative and still too narrow(89).

Two extension approaches have been proposed to overcome the problem of imprecision of estimated variance between trials from a frequentist perspective. One approach, proposed by Hardy and Thompson(89), uses the profile likelihood intervals method to allow for asymmetric intervals and uncertainty in the estimate of τ^2 for further estimation of overall treatment effect. The approach also provides information on the confidence interval of τ^2. Thus, the approach yields a wider confidence interval than the standard random effects approaches. The approach is suggested to be used instead of the standard methods in random effects meta-analysis when the value of τ^2 has a merit impact on the overall estimated treatment effect. Details of the method are provided in the paper by Hardy and Thompson (89).

Another approach is proposed by Biggerstaff and Tweedie (88). They also attempt to take into account variation of the point estimate of τ^2 of DerSimonian and Laird in estimating the overall treatment effect. They propose a new method to calculate the weights given to individual trials. One benefit of the new method is to obtain an approximate distribution of DerSimonian and Laird τ^2 from developing a simple form for the variance of Q statistic. Another way is by obtaining the distribution of DerSimonian and Laird τ^2 from asymptotic likelihood methods. Details of the method are provided in the Biggerstaff and Tweedie paper (88). This approach produces down-weighting of the results for large studies and up-weighting of the results for small studies. The authors discuss that their method will give different results from those of the standard random effects model when the number of trials is fairly small (< 20 based on the results of Larholt, et al(110)).

5.5.3 Fixed effect model versus random effects model

Although the assumptions of fixed and random effects models are clearly different, in practice it is still difficult to decide on an appropriate model for combining the trial results of interest, especially when the homogeneity testing result is marginal. In fact, none of the two approaches is considered to be the perfect model for all meta-analyses, especially when small number of trials are included in the meta-analyses.This is, however, common in meta-analyses. Many comments and disputes on the selection of an appropriate model have been discussed(85). Nevertheless, the random effects approaches always give a wider confidence interval of the overall treatment effect than the fixed effect approaches when the between trial variance is larger than zero($\tau^2>0$). Currently, several authors(7, 15, 90, 111) suggest to use the random effects model routinely, since similar results to the fixed effect model will be obtained when $\tau^2 \cong 0$. In addition, the results can be used to evaluate robustness of the results obtained from fixed effect model.

Topic 6: General linear mixed models (GLM)

6.1 Introduction

General linear mixed models (GLM) are the statistical approaches incorporating both fixed- and random-effects terms allowing for heterogeneity between trials in the effect of treatment of interest(109). The approaches are often called mixed effect models, as suggested by Hedges in Chang et al (112). The GLM can be used to detect and explain heterogeneity in meta-analyses. Treatment effect of individual trials, e.g. log-relative risk, is treated as a response variable, which is related to potential factors, called covariates, from the same trials and some random effects terms. In some situations, variability between trials within each category of some categorical covariates need to be considered. Here, the GLM can be extended to incoporate random components of individual categories of the covariates. This extension is also illustrated in the application section.

The covariates may be some potential factors at trial level, such as trial designs, treatment schedule, outcome measures, etc, and at subject level such as different gender proportions, age of subjects and follow-up period. The data at subject level is usually in aggregated form, e.g. subject mean age, mean follow-up period, etc. This is because most of the data analysed in meta-analysis is extracted from completed trials unless individual subject data on each trial could be gathered from the authors. However, the latter situation is rather difficult to achieve and is hardly possible in most practices.

In this topic, section 6.2 discusses setting of GLM for meta-analyses related to cluster randomized trials when the binary outcome is measured in log-relative risk. Section 6.3 describes the assumption of the model. Section 6.4 discusses the approach used to estimate parameters of the model. Section 6.5 discusses the estimation of confidence intervals for parameters of the model. Section 6.6 provides the application of GLM to the three published meta-analyses. Finally, section 6.7 gives a summary of the application of GLM on the meta-analyses.

6.2 The model

The simple two-level variance components model that allows for within-study variation at level-one and between-study variation at level-two is used to pool the treatment effects from individual trials. In the model, log-relative risk obtained from each trial i, θ_i, is treated as a continuous dependent variable. The randomization design is treated as a binary covariate X at trial level. X equals 1 for cluster randomized design and equals 0 for individually randomized design. Adding other potential covariates can extend the model. The model can be expressed (113) as

$$\theta_i = \theta + \beta_1 x_{i1} + \sum_{j=2}^{p} \beta_j x_{ij} + u_i + \varepsilon_i \qquad (6.1)$$

where θ is a fixed coefficient that represents an overall treatment effect and

x_{i1} is the randomization design covariate for trial i.

$x_{i2},..,x_{ip}$ are the values of the j known potential covariates, j=2,…,p, for trial i.

$\beta_1,....,\beta_p$ are the fixed coefficients indicating association between its related covariate and outcome of treatment effect.

u_i is an unobserved random effects term, which represents the deviation of trial i specific-mean $\bar{\theta}_i$ from the overall effect θ adjusted for the effect of covariates, $\sum_{j=1}^{p} \beta_j x_{ij}$.

ε_i is the random effects term that represents sampling error of trial i.

6.3 Assumptions

In the parametric approach the random effects are usually assumed to be independent normal distributions. For the simple GLM involving continuous response θ_i, random effects in the sampling errors ε_i is assumed normal with mean zero and variance σ_i^2.

In the level-two model, the unobserved random effects u_i is also assumed to be normally distributed with mean zero and variance τ^2.

The level-one random error ε_i and level-two unobserved random effects u_i are assumed to be independent, consequently

$$\mathrm{cov}(u_i, \varepsilon_i) = 0$$

and

$$E(\theta_i) = \theta + \sum_{j=1}^{p} \beta_j x_{ij} \ , \ \mathrm{var}(\theta_i) = \sigma_i^2 + \tau^2.$$

To fit the model (6.1) to the data, θ, β_j and the variances σ_i^2, τ^2 are estimated. In practice the estimated variance of log-relative risk $\hat{\sigma}_i^2$ is available from individual trial i. It is usually used as an estimate for σ_i^2.

6.4 Parameter estimation

Under the assumption of multivariate normal distribution of random effects, the model parameters are estimated by restricted maximum likelihood(REML). For computation procedure, the restricted iterative generalized least squares (RIGLS) algorithm is used via the MLwiN software(114). Even when standard errors for variance estimates are provided, they are based on asymptotic properties and may be unreliable except in very large samples. Thus, it might be better to make inferences based on bootstrapped standard errors (115).

6.5 Confidence interval calculation

The parametric bootstrap estimation is used in this study. The estimation requires no normality assumption for the estimate from which the confidence interval is calculated. Thus, it is useful when the sample sizes are small, especially for the variance τ^2. Furthermore, ranges obtained from the method are interpreted as approximate confidence interval for θ and τ^2 with relaxing the normality assumption strongly required in the likelihood method (116).

The bootstrap confidence intervals for the true overall treatment effect θ and the variances τ^2 are generated from 1000 replications using the MLwiN software(117). The 2.5 and 97.5 percentiles are used as a range of the 95 per cent confidence for the parameters. The lower limit of confidence interval for the variance τ^2 will always be zero when the 2.5 percentile of the bootstrap distribution is in the negative value. This is because a negative variance cannot be interpreted.

6.6 Application to published meta-analyses

In this section, the applications of GLM to the meta-analyses related to cluster randomized trials are illustrated in the three different examples. The analysis is implemented using the MLwiN software.

6.6.1 Meta-analysis of vitamin A supplementation trials

The observed log-relative risks of child mortality for individual trials are considered as a continuous response variable. The model (6.1) is fitted with only the intercept representing an overall treatment effect, and random effects components of level-one and level-two as

$$\theta_i = \theta + u_i + \varepsilon_i \qquad \text{(a)}$$

Here, the main parameters of interest from the model are θ and τ^2. They are presented in table 6.1.

Table 6.1 Estimated parameters (95 per cent CI) for model (a) on log scale

Model	Estimated parameters		-2log-likelihood
	Fixed effect	Random effect	
	$\hat{\theta}$	$\hat{\tau}^2$	
Fixed effect	-0.31 (-0.40, -0.22)	-	10.01
(a)	-0.36 (-0.60, -0.15)	0.08 (0.00, 0.15)	7.55

The GLM produces an estimated effect of vitamin A supplementation compared to the control group as a log-relative risk of -0.36 (95 per cent CI -0.60 to -0.15). The figure gives the significant protective effect of vitamin A supplementation. The estimated variance of between trials is 0.08 (95 per cent CI 0.00 to 0.15). The likelihood ratio obtained from the regression model can be used to test heterogeneity effect. The likelihood ratio statistic is calculated from a difference between –2log-likelihood of fixed effect model, 10.01 and model (a), 7.55. The square root of the ratio of 2.46 gives p = 0.058 on the asymptotic distribution of the likelihood ratio statistic. This figure shows the marginal evidence of heterogeneity of the treatment effect across trials.

The results show evidence of the beneficial effect of vitamin A supplementation to reduce child mortality and non-significant variability of treatment effect between trials. Some other factors affecting the log-relative risk may exist but these are not available for investigation in the model such as various units of treatment allocation and different control groups across trials.

6.6.2 Meta-analysis of mammographic screening trials

The meta analysis is performed to evaluate effect of mammographic screening on reduction of breast cancer mortality in women aged less than 50 years. It includes eight identified trials perfomed in many western countries. The primary outcome is breast cancer mortality. Log-relative risks for individual trials are used as a continuous response variable. The steps of fitting data to the model (6.1) are similar to those in the previous example. The model is expressed as

$$\theta_i = \theta + u_i + \varepsilon_i \qquad \text{(b)}$$

The design variable is then added to the model (b) as a fixed effect. The extended model is shown as

$$\theta_i = \theta + \beta_1 \text{Design}_i + u_i + \varepsilon_i \qquad \text{(c)}$$

The model (c) is extended to consider whether any difference in heterogeneity exists between trials of each group according to the randomization design. This allows the effect of

randomization design to vary between trials. Defining $u_{i1} \sim N(0, \tau^2_{CRT})$ and $u_{i2} \sim N(0, \tau^2_{IRT})$ as the independent random effects, the model is then written as

$$\theta_i = \theta + \beta_1 Design_i + u_{i1} Design_i + u_{i2}(1 - Design_i) + \varepsilon_i \qquad (d)$$

Here τ^2_{CRT} represents the variance between trials of cluster randomized design group and τ^2_{IRT} represents the variance between trials of individually randomized design group. They are presented in table 6.2.

Table 6.2 Estimated parameters (95 per cent CI) for model (b) to (d) on log scale

Model	Estimated parameters					-2log-likelihood
	Fixed effect		Random effects			
	$\hat{\theta}$	$\hat{\beta}_1$ (Design)	$\hat{\tau}^2$	$\hat{\tau}^2_{CRT}$	$\hat{\tau}^2_{IRT}$	
Fixed effect	-0.22 (-0.32, -0.13)	-	-	-	-	-1.03
(b)	- 0.23 (-0.40, -0.10)	-	0.02 (0, 0.04)	-	-	-2.60
(c)	-0.17 (-0.45, 0.10)	-0.10 (-0.36, 0.17)	0.03 (0, 0.05)	-	-	-2.91
(d)	-0.19 (-0.55, 0.15)	-0.08 (-0.44, 0.27)	-	0.02 (0, 0.03)	0.06 (0,0.10)	-2.88

The model (b) produces an estimated log- relative risk of -0.23(95 per cent CI -0.40 to -0.10). The result shows the significant protective effect of mammographic screening, compared to the control group. The estimated variance of random effects is found to be 0.02 (95 per cent CI 0.00 to 0.04). It reflects a slight heterogeneity between trials. The likelihood ratio statistic is 1.57 (-1.03-(-2.60)) and its square root gives p = 0.11 on the asymptotic distribution of the likelihood ratio statistic.

When the covariate of randomization design is added to the model as in (c), its effect on log-relative risk is small and non-significant. In addition, the estimated variance of random effects is about the same, which is as small as the variance of model (b). These figures may explain the slight change in estimate of the adjusted overall log-relative risk if compared to model (b). There is inconclusive benefit of the mammographic screening on breast cancer mortality after the randomization design has been adjusted for.

When allowing for random effects to log-relative risk in the two randomization design groups as in model (d), the estimated variance of between trials for individually randomized design group is larger than that of the cluster randomized design group. However, both are in small values. The result may reflect a more similar effect of mammographic screening effect on breast cancer mortality in the group of cluster randomized trials. Here, the estimate of adjusted effect of mammographic screening in model (d) is –0.19 (95 per cent CI –0.55 to 0.15). It does not differ from the results of model (c).

The results show inconclusive evidence on the benefits of mammographic screening on breast cancer mortality in women aged less than 50 years, after the randomization design is adjusted.

6.6.3 Meta-analysis of multiple interventions trials

The original meta-analysis is done to assess the effectiveness of multiple risk factor interventions to reduce cardiovascular risk factors from coronary heart disease. Study subjects are adults aged at least 40 years and having no clinical evidence of established cardiovascular disease.

Log-relative risks of smoking for individual trials are fitted to the GLM as a continuous response. Sequences of fitting data to the model (6.1) follow the previous example. The estimates and 95 per cent confidence intervals of the parameters are presented in table 6.3.

Table 6.3 Estimated parameters (95 per cent CI) for model (e) to (g) on log scale

Model	Estimated parameters					-2log-likelihood
	Fixed effect		Random effect			
	$\hat{\theta}$	$\hat{\beta}_1$ (Design)	$\hat{\tau}^2$	$\hat{\tau}^2_{CRT}$	$\hat{\tau}^2_{IRT}$	
Fixed effect	-0.15 (-0.18, -0.12)	-	-	-	-	25.38
(e)	-0.11 (-0.18, -0.04)	-	0.01 (0, 0.02)	-	-	-20.56
(f)	-0.12 (-0.21, -0.03)	0.03 (-0.10, 0.15)	0.01 (0, 0.02)	-	-	-20.68
(g)	-0.12 (-0.20,-0.02)	0.03 (-0.09,0.14)	-	0.005 (0,0.01)	0.013 (0,0.03)	-21.56

Model (e), which has no covariate, yields the estimated overall log-relative risk of -0.11 (95 per cent CI -0.18 to -0.04). This figure shows the significant protective effect of multiple interventions compared to the control. The estimated variance of treatment effect between trials is 0.01(95 per cent CI 0.00 to 0.02), which gives light variability in the effect of intervention from trial to trial. The likelihood ratio statistic is 45.97 (25.38-(-20.56)) and its square root gives p = 6.00e-12 on the asymptotic distribution of the likelihood ratio statistic.

The effect of the randomization design in model (f) is very small, with a log-relative risk of 0.03 (95 per cent CI -0.10 to 0.15). The estimate of adjusted overall log-relative risk of multiple interventions is quite similar to the unadjusted results in model (e). The estimate of adjusted variance of between trial log-relative risks is also similar to the unadjusted results in model (e).

When two random components for the randomization design are incorporated as in model (g), the adjusted effects of multiple interventions and randomization design remain similar to the results in model (f). The estimate of adjusted variance of between trial log-relative risk for the group of individual randomized design, 0.013 (95 per cent CI 0, 0.03), is about two times larger than that for the group of cluster randomised design, 0.005 (95 per cent CI 0.00, 0.01). However, these variances still remain very small.

Table 6.3 shows that the models (e) to (g) provide quite similar log-likelihood results. This information is relevant to the finding of similar estimated log-relative risk of treatment effect and other covariates.

The results from this example show that adjustment of randomization design hardly affects the overall intervention effect. The variability of treatment effect between trials is also very small.

6.7 Summary

The GLM shows that putting the potential covariates and some random effects components to the model easily provides investigation of heterogeneity between trials. More information is also given to explain heterogeneity between trials due to covariates effects and variability of random effects, compared to the simple conventional methods. Despite these advantages, care should be taken when using the model, as it needs a strong assumption of normality distribution of random effects components. It is also difficult to verify validity of the assumption.

Topic7: Generalized linear mixed model (GLMM)

7.1 Introduction

Heterogeneity or variability in trial results is common in meta-analysis. Several researchers (7, 109, 111, 118-121) have proposed GLMM to fit the meta-analysis data to investigate and explain heterogeneity. Available potential factors at the trial and subject levels in the reviewed papers and unobserved random effects are included to the models to explain heterogeneity. The normality assumption of unobserved random effects is usually adopted to estimate the model parameters. There are some concerns on the validity of such an assumption. This may, therefore, lead to misleading conclusions from the model.

Nonparametric maximum likelihood estimator (NPML) is an alternative approach proposed by some authors (119, 122-124)for approximate estimation in GLMM. The NPML is estimated from discrete mixing distributions. Aitkin(122) and Dietz (118) have shown the approach in some problems of the meta-analyses. This chapter discusses the GLMM with NPML in particular emphasis on meta-analyses involving cluster randomized trials. The performance of NPML is examined to solve the problems of heterogeneity of treatment effect between trials.

In the previous topics, observed log-relative risk, which is a common summary measure for a binary outcome in meta-analysis, is used as a response variable. Here, to investigate more information of random treatment effects, observed number of events for each treatment group of individual trials is used as a response variable and treatment groups are treated as a covariate of a model.

Section 7.2 discusses the model setting for meta-analysis involving cluster randomized trials with two treatment groups measured in binary outcome. Section 7.3 presents the NPML to estimate parameters of the models. Section 7.4 discusses classification of trials to explain heterogeneity between trials. Section 7.5 illustrates application of the approach to the three meta-analyses published in the literature. Section 7.6 presents a summary on the GLMM in the meta-analyses studied.

7.2 The model for meta-analysis involving CRTs

To fit the GLMM to meta-analysis involving cluster randomized trials with binary outcome, the simple two-level variance components is used. The model allows for within-trial variation at level-one and between-trial variation at level-two.

Here the observed number of events y_{it}, in the treatment group t, t=1,2, with a sample size n_{it} from individual trials i, i=1,...I, is a response variable. The observed y_{it} may have a binomial distribution,

$$Y_{it} \sim \text{Binomial}(p_{it}, n_{it})$$

or poisson distribution, $Y_{it} \sim \text{Poisson}(p_{it})$

where p_{it} is the mean of individual events of treatment group t in trial i. The mean p_{it} is associated with linear predictors through a canonical link function,

$$g(p_{it}) = LP_{it}$$

From the meta-analysis studied, effects of treatment, randomization design and other potential factors are treated as covariates of the model, thus

$$LP_{it} = \theta \text{treat}_t + \delta x_{it} + \gamma \text{design}_i + \beta x_i + u_i \qquad (7.1)$$

where treat_t is the binary variable of treatments t that assigns 1 for treatment and 0 for control.

design_i is the binary variable of randomization design of trial i that assigns 1 for cluster randomized trials and 0 for individually randomized trials.

x_{it} is a vector of other potential factors of subjects in treatment group t for trial i, such as mean blood pressure of subjects in treatment and control groups.

x_i is a vector of other potential factors or may be the interaction effect between randomization design and treatment effect.

θ is an unknown fixed effect of the treatment on log scale.

δ, γ, β are unknown fixed effect parameter vectors on log scale.

u_i is random effects of between trials. It has a distribution $\phi(u)$ that remains unspecified.

For the binomial model, the canonical link function $g(p_{it}) = \log(p_{it}/(1-p_{it}))$ thus model (7.1) can be replaced to be

$$\log(p_{it}/(1-p_{it})) = \theta treat_t + \delta x_{it} + \gamma design_i + \beta x_i + u_i \qquad (7.2)$$

The estimate of adjusted odds ratio of treatment effect is easily calculated from the exponential of θ.

When the trials have different follow-up periods, the poisson model is an appropriate model to fit this kind of data because it takes into account the follow-up period for each trial by adding the offset term of log transformation of person-time to the model. For the poisson model, the canonical link function is $g(p_{it}) = \log(p_{it})$. The model (7.1) can be rewritten as

$$\log(p_{it}) = \theta treat_t + \delta x_{it} + \gamma design_i + \beta x_i + u_i + \log(person-time)_i \qquad (7.3)$$

Here, the estimate of adjusted relative risk is also the exponential of θ.

The fixed treatment effect can extend to be a random treatment effects to evaluate treatment heterogeneity. To achieve this purpose, $\theta treat_t$ is replaced by the term $(\theta + z_i)treat_t$ in model (7.1). It becomes

$$g(p_{it}) = LP_{it} = \theta treat_t + \delta x_{it} + \gamma design_i + \beta x_i + z_i treat_t + u_i \qquad (7.4)$$

where z_i is a random effect of the treatment. Thus u_i and z_i have an unknown joint distribution $\phi(u, z)$.

7.3 Nonparametric maximum likelihood (NPML) estimator of parameters

At this step the likelihood function is determined to obtain maximum likelihood estimates of the model parameters. When the treatment is fitted to the model as a fixed effect, like model (7.1), the likelihood function is then

$$L(\theta, \delta, \gamma, \beta, \lambda) = \prod_{i=1}^{I} \int \prod_{t=1}^{2} f(y_{it}|\theta, \delta, \beta, \gamma, u_i)\phi(u_i)du_i \qquad (7.5)$$

where λ is a parameter vector of ϕ.

The function $f(y_{it}|\theta, \delta, \beta, \gamma, u_i) = f(y_{it}|LP_{it})$ denotes the probability density for y_{it} given the linear predictors. Since the distribution $\phi(u)$ is not known, the model parameters are estimated non-parametrically. Here, it is reasonable to consider the $\phi(u)$ as a discrete distribution with K components, when $K \le I$ (118). Aitkin and Dietz (118, 119, 122) discuss this point and show that

$$\phi(u) = \pi_k \text{ for all } (u) = (u_k)$$
$$= 0 \text{ or else}$$

might give a good approximation for the free distribution $\phi(u)$ if u and π are chosen approximately.

In the distribution $\phi(u)$, $u = (u_1, \ldots, u_K)$ and $\pi = (\pi_1, \ldots, \pi_{K-1})$ are the parameters, and $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$. π_k denotes the mixture proportion at known mass point u_k.

The likelihood function (7.5) can now replace the integral over u_i by a finite sum of K components. Some authors (119, 123, 125, 126) discuss this concept to be due to the integral not having a closed form except for the normal distribution of the response variable. The likelihood then becomes

$$L(\theta,\delta,\gamma,\beta,\pi,u) = \prod_{i=1}^{I}\sum_{k=1}^{K}\pi_k\prod_{t=1}^{2} f(y_{it}|LP_{itk}) \qquad (7.6)$$

The log-likelihood function is as follows:

$$LL(\theta,\delta,\gamma,\beta,\pi,u) = \sum_{i=1}^{I}\log\sum_{k=1}^{K}\pi_k\prod_{t=1}^{2} f(y_{it}|LP_{itk}) \qquad (7.7)$$

Therefore, the functions (7.6) and (7.7) are respectively the likelihood and log likelihood of a finite mixture GLMM with known proportion π_k at known mass point u_k, with the linear predictor for the trial i in the mixture component k being (123)

$$LP_{itk} = \theta treat_t + \delta x_{it} + \gamma design_i + \beta x_i + u_k \qquad (7.8)$$

Although the number K component is an unknown parameter, it is treated as fixed, and successively increased until the likelihood is maximized (127). If K is the number of components that maximizes the likelihood, then $\hat{\pi}$ and $\hat{u}$ are the NPML estimates of ϕ. The maximum likelihood of such interest models is implemented by the EM-algorithm procedure, discussed in Dietz and Böhning 1994(128), Dietz and Böhning 1995(129) and Aitkin(119). Implementation is done using the PORML macro programme in S-plus software. This macro programme is supported by Dr Dietz (130). The optimal K component is then selected by using the Bayesian Information Criterion (BIC). The BIC is estimated (131, 132) as

$$BIC = 2LL(\hat{\theta},\hat{\delta},\hat{\gamma},\hat{\beta},\hat{\pi},\hat{u}) - P\log(I) \qquad (7.9)$$

where $LL(\hat{\theta},\hat{\delta},\hat{\gamma},\hat{\beta},\hat{\pi},\hat{u})$ is the log-likelihood of estimates for K component.

P is the number of parameters estimated freely. For example; if we have 2 components, there must be P=3. This number comes from 2 component-specific means of treatment effect (θ_1 and θ_2) plus one component weight (p_1) to be estimated, as p_2 is equal to $(1-p_1)$.

I is the number of trials in meta-analysis.

The K component that gives the largest value of BIC is the optimal K.

When adding treatment effect as a random effect, the nonparametric maximum likelihood estimates of parameters are obtained from the model in the same way as for the fixed treatment effect model. First, the likelihood function to be maximized is summed over K components under the discrete join distribution $\phi(u,z)$,

$$L(\theta,\delta,\gamma,\beta,\pi,u,z) = \prod_{i=1}^{I}\sum_{k=1}^{K}\pi_k\prod_{t=1}^{2} f(y_{it}|LP_{itk}) \qquad (7.10)$$

and then the log-likelihood is

$$LL(\theta,\delta,\gamma,\beta,\pi,u,z) = \sum_{i=1}^{I}\log\sum_{k=1}^{K}\pi_k\prod_{t=1}^{2} f(y_{it}|LP_{itk}) \qquad (7.11)$$

Function (7.10) and (7.11) are likelihood and log likelihood of a finite mixture GLMM with known proportion π_k at known mass point (u_k, z_k), with the linear predictor for the trial i in the mixture component k being(123)

$$LP_{itk} = \theta treat_t + \delta x_{it} + \gamma design_i + \beta x_i + u_k + z_k treat_t \qquad (7.12)$$

The maximum likelihood estimates of u, z, π and other fixed effect parameters at the K^{th} component are obtained by using the same procedure as in the previous discussion. However, to implement model (7.12) in the S-plus, an interaction term between u_k and $treat_t$ is needed.

At this stage the estimates u, z, π obtained from the model can be used to calculate a weighted average treatment effect as

$$\hat{\theta} = \sum_{k=1}^{K} \hat{\pi}_k \hat{z}_k \tag{7.13}$$

and the variance of the mixing distribution on z_k

as $$Var(\hat{\theta}) = \sum_{k=1}^{K} \hat{\pi}_k \hat{z}_k^2 - (\sum_{k=1}^{K} \hat{\pi}_k \hat{z}_k)^2 \tag{7.14}$$

This information can provide an explanation for treatment heterogeneity. If the $\hat{\theta}$ obtained from (7.13) is similar to the results obtained from model (7.1) and $Var(\hat{\theta})$ approaches zero, no evidence of variability between K components of trials is shown. The estimate of fixed treatment effect may be the appropriate answer for treatment effect.

In addition, a variance of the baseline heterogeneity among the K components is estimated as

$$\hat{\tau}^2 = \sum_{k=1}^{K} \hat{\pi}_k \hat{u}_k^2 - (\sum_{k=1}^{K} \hat{\pi}_k \hat{u}_k)^2 \tag{7.15}$$

This value is expressed as the variability between K components beyond the treatment effect. It can be used as an indicator of heterogeneity of unobserved random effects.

Many authors (118, 119, 122, 123, 133) have pointed out that interpretation of estimate treatment effect under the discrete distribution of random effects is rather difficult since the exact distribution of the random effects is unknown. The results obtained from the finite mixture distribution of the NPML approach could, however, provide reasonable information on heterogeneity. If only one component in a mixture distribution is obtained, no heterogeneity across trials is considered.

7.4 Classification of trials

When the maximum likelihood estimates of parameters are obtained at the K^{th} mixture component where $K > 1$, the trials can be classified to each component of the mixture distribution. Classification is achieved by using the maximum posterior probability that the trial I comes from the K^{th} component (118, 127). The maximum posterior probability is expressed(118) as

$$pr(trial_i \in C_k \mid y_{1i}, y_{2i}, \hat{\pi}, \hat{u}, \hat{z}, \hat{\theta}, \hat{\gamma}, \hat{\beta}) = \frac{\hat{\pi}_k \prod_{t=1}^{2} f(y_{it} \mid \hat{LP}_{itk})}{\sum_{r=1}^{K} \hat{\pi}_r \prod_{t=1}^{2} f(y_{it} \mid \hat{LP}_{itr})} \tag{7.16}$$

where C_k is the k^{th} component and

$\hat{LP}_{itk}$ is the estimated linear predictor for trial i in the component k of the mixture distribution.

The results obtained from the classification can be used to explain the finding of heterogeneity that may be beyond the effect found in the model. This is a superior point of the NPML estimator as compared to the parametric maximum likelihood estimator.

7.5 Applications to published meta-analyses

In this section, the GLMM via NPML estimator is illustrated in the three published meta-analyses described in topic 4. Because each of these meta-analyses includes trials with different follow-up periods, the analysis is therefore performed by using the maximum likelihood estimation of a 2-level mixed poisson regression models, PORML macro programme in S-plus software. To account for mean follow-up periods from individual trials, the log transformation of multiplication of sample size and mean follow-up period must be taken as an OFFSET for the model (134).

7.5.1 Meta-analysis of vitamin A supplementation trials

This meta analysis(1) includes eight community-based trials to determine the relationship of vitamin A supplementation and mortality in children aged 6 to 72 months. To fit a poisson regression model to the data, the observed number of child deaths from each treatment group of individual trials is the response variable.

The analysis of mixed poisson regression models for this data set is performed in two steps. First, the baseline heterogeneity from the random effects term of the model is investigated. The model is fitted with a fixed effect of treatment and a random intercept term, ignoring any covariate, as model A in table 7.1. The mixing distribution is estimated non-parametrically starting from K=2 mixture components. The number of component is increased systematically until the deviance is stable and it gives the largest BIC value. The non-parametric maximum likelihood estimates are obtained at K = 5 with the BIC of 1451.24. The model produces an estimate of the fixed effect of vitamin A supplementation as log-relative risk of –0.31(95 per cent CI -0.45 to -0.17). The variance of mixing distribution is 0.93. It represents a huge variation of the baseline between mixture components on log scale.

In the second step, the treatment effect is included as a random part of the model to evaluate the treatment effect heterogeneity, as model B in table7.1. This model is called a full random slope and intercept model. The maximum likelihood of a finite mixture model is also estimated by the NPML. At K= 5 again the maximum likelihood estimate is obtained with the BIC value of 1463.65. The estimate of weighted average log-relative risk of treatment effect is –0.43 (95 per cent CI –1.37 to 0.51). At this step even the estimate of the average effect of vitamin A supplementation is more effective, the confidence interval is much wider than the result obtained from model A. The upper limit of confidence interval is also higher than zero. This figure represents evidence of random treatment effects. The variance of mixture distribution for the baseline remains at a very high value of 0.89.

Table 7.1 NPML estimates of treatment effect and variance of random effects for each model in meta-analysis of vitamin A supplementation trials

Model	K Components	BIC	Baseline variance ($\hat{\tau}^2$)	Overall treatment effect	
				Estimated log-relative risk ($\hat{\theta}$)	95 per cent CI for θ
A	5	1451.24	0.93	-0.31	-0.45, -0.17
B	5	1463.65	0.89	-0.43	-1.37, 0.51

A = k components mixture distribution of baseline effect and fixed treatment effect
B = k components mixture distribution of baseline effect and treatment effect

The model produces further results of the classification of trials to K components of the mixture distribution. This information is used to further explain the results of treatment effect and baseline heterogeneity. Table 7.2 presents a wide range of component- specific log-relative risk from -1.60 to 0.03. Three, at k =3,4,5, of the five components show evidence of relative similar log-relative risk in beneficial effect of treatment. The component-specific log-relative risk of the 2nd component presents increasing risk, but one that is inconclusive. Some explanation can be given for these results.

The 1st component shows a very high effect of the treatment, determined by trial 5. The trial has a small sample size and very small number of child deaths in the vitamin A supplementation group. In addition, this trial has a mean follow-up period of 42 months, which is a much longer period than the other trials of around 5-18 months. The 4th component is determined by trial 7 having substantial different units of treatment allocation, wards, compared to the other trials. The 2nd component is determined by trial 3 and trial 8, which have similar inconclusive treatment effects. The 3rd and 5th components are determined by trials 1, 4, 2, and 6 respectively. These trials have likely similar units of treatment allocation and the follow-up periods.

Table 7.2 NPML estimates of treatment effect distribution and classification of trials according to the 5 components of the maximum likelihood estimates for model B

Kth component	Estimated log-relative risk (95 per cent CI)	Weight	Trial number
1	-1.60 (-2.76, -0.44)	0.125	5
2	0.03 (-0.28, 0.34)	0.250	3, 8
3	-0.46 (-0.77, -0.16)	0.250	1, 4
4	-0.35 (-0.64, -0.05)	0.125	7
5	-0.32 (-0.53, -0.12)	0.250	2, 6

For this example, the GLMM via NPML approach provides evidence of inconclusive effect of vitamin A supplemention on child mortality with a wide confidence interval of the true treatment effect. It also presents heterogeneity of treatment effect. The result obtained from this study does not correspond to the result from the original meta-analysis paper. The model also gives a huge variability of baseline characteristics. One point to note is that this meta-analysis has a small number of trials.

7.5.2 Meta-analysis of mammographic screening trials

This meta-analysis (2) is performed to evaluate the effect of mammographic screening on reduction of breast cancer mortality in women aged less than 50 years. It includes eight identified trials perfomed in many western countries. The number of woman deaths with breast cancer from individual trials are treated as a response variable of the poisson regression model.

Analysis is similar to the steps in the previous example. The first step is to investigate baseline heterogeneity. The model is fitted with screening programme as a fixed effect and a random intercept term, ignoring covariates, as model C in table 7.3. The NPML approach is used to produce maximum likelihood estimates of the model parameters with successive mixture components from K= 2. Estimating the mixing distribution non-parametrically gives the maximum likelihood estimates at K= 4 with the BIC of 146.00. Estimated variance of the

mixing distribution is 0.10. This reflects modest baseline heterogeneity on log scale. The estimated log-relative risk of the fixed treatment effect is –0.21(95 per cent CI -0.35 to-0.07).

In the second step, the heterogeneity effect of screening programme is investigated. Here it is fitted into the model as a random effects, as model D in table 7.3. The NPML gives the maximum likelihood estimate at 4 mixture components with the BIC of 145.37. This BIC value is similar to the result obtained from model C. Estimated variance of the baseline mixture distribution adjusted for the random treatment effects is 0.09, which is similar to the result of model C. The estimate of weighted average log-relative risk is -0.25 (95 per cent CI-0.58 to 0.08). The confidence interval here is much wider than the result from model C. This figure represents some evidence of random treatment effects.

The third step is to investigate heterogeneity of some available covariates. In this meta-analysis, some trials randomly allocated the screening programme to groups of women rather than to individual women. This difference of treatment randomization design is considered as a binary variable: 1 for group randomized and 0 for individual randomized, in the mixed poisson regression model. The randomization design is added to the model D as a fixed effect, and now in model E in table 7.3. The NPML gives the maximum likelihood estimate at K=3 with the BIC of 142.46. The randomization design gives the non-significant effect with an estimated log-relative risk of 0.12 (95 per cent CI -0.04 to 0.28). Estimated variance of the mixing treatment distribution is 0.07 on log scale. It is slightly different from the results of previous models. The estimate of adjusted effect of screening programme on breast cancer mortality is the log-relative risk of -0.23 (95 percent CI -0.35,-0.11). This adjustment illustrates similar effects of screening programme to the results obtained from the fixed treatment effect in model C.

In addition, the interaction terms of (design*treat) is added to the model for further investigation. The effect is not significant and the model gives similar results of the treatment effect to model E. The model including an interaction term is not presented .

Table 7.3 NPML estimates of treatment effect and variance of random effects for each model in the meta-analysis of mammographic screening trials

Model	K Components	BIC	Baseline variance ($\hat{\tau}^2$)	Overall treatment effect	
				Estimated log-relative risk ($\hat{\theta}$)	95 per cent CI for θ
C	4	146.00	0.10	-0.21	-0.35, -0.07
D	4	145.37	0.09	-0.25	-0.58, 0.08
E	3	142.46	0.07	-0.23	-0.35, -0.11

C = k components mixture distribution of baseline effect and fixed treatment effect
D = k components mixture distribution of baseline effect and treatment effect
E = k components mixture distribution of baseline effect and treatment effect plus fixed randomization design

The classification of trials according to model E, which adjusts for randomization design, is presented in table 7.4. The table shows various component-specific log-relative risks from –0.33 to –0.08. Confidence intervals for the component-specific log-relative risk in the 1st component and 3rd component are very wide and inconclusive. The 1st component is determined by trial 6, with very imbalanced treatment groups. The 3rd component is determined by trial 8, with practices as units of treatment allocation, where it differs from the

rest. Most of the trials belong to the 2nd component, where the results of mammographic screening are found to be significant beneficial effect.

The results illustrate some heterogeneity effects of mammographic screening and moderate baseline heterogeneity. These results may indicate that the estimate of overall log-relative risk may not appropriately represent the treatment effect for this example, even when there is little baseline variability.

Table 7.4 NPML estimates of treatment effect distribution and classification of trials according to the 3 components of the maximum likelihood estimates of model E

K^{th} Component	Estimated log-relative risk (95 per cent CI)	Weight	Trial number
1	-0.33 (-0.86, 0.20)	0.130	6
2	-0.23 (-0.38, -0.08)	0.745	1-5, 7
3	0.08 (-0.39, 0.22)	0.125	8

7.5.3 Meta-analysis of multiple interventions trials

This meta-analysis (3) is done to assess the effectiveness of multiple risk factor interventions to reduce cardiovascular risk factors from coronary heart disease. Study subjects are adults aged at least 40 years and having no clinical evidence of established cardiovascular disease. In this thesis, reanalysis is performed for the fourteen trials providing the outcome of smoking prevalence.

First, baseline heterogeneity is evaluated by fitting the mixed regression model with a fixed intervention effect and a random intercept term, ignoring covariates, as model F in table 7.5. The NPML approach with sequentially mixture components from K=2 is used to obtain the maximum likelihood estimate from the model. The NPML approach gives the maximum likelihood estimate at 4 mixture components with the BIC of 2692.91. The variance of the mixing distribution is 0.29 on log scale. This figure shows considerable baseline heterogeneity. The estimate effect of multiple interventions on smoking prevalence is the log-relative risk of -0.16 (95 per cent CI-0.20 to-0.12), which gives the significant protective effect of the multiple interventions.

Next, the fixed effect of multiple interventions is replaced by a random effects as in model G. This is performed to evaluate the heterogeneity effect of multiple interventions. The NPML again gives the maximum likelihood estimate at K=4 with the BIC of 2737.61. The estimated variance of mixing distribution of baseline is 0.31, representing similar variation to the previous model results. The estimate of weighted average log-relative risk of multiple interventions is -0.11 (95 per cent CI-0.25, 0.03), which is relatively similar to the result of model F. Here, the result is inconclusive.

In the third step, further heterogeneity is investigated by adding the covariate of randomization design as a fixed effect to the model G, becoming the model H. The NPML produces the maximum likelihood estimate again at K=4 with the BIC of 2740.23. The effect of randomization design is not significant. The results of estimated variance of mixing distribution of baseline and estimate of adjusted weighted average log-relative risk of the multiple interventions are still similar to the result of model G.

Next, the interaction between treatment and randomized design is fitted to the model H, now becoming model I. A significant effect of the interaction is seen. Estimates of the parameters of this model are presented in table 7.6. Here, a slight raise in the estimate of adjusted weighted average log-relative risk of the multiple interventions is seen in protective effect with a significant result of -0.19 (95 per cent CI-0.35, -0.03).

Table7.5 NPML estimates of treatment effect and variance of random effects for each model in the meta-analysis of multiple interventions trials

Model	K Components	BIC	Baseline variance($\hat{\tau}^2$)	Overall treatment effect	
				Estimated log-relative risk ($\hat{\theta}$)	95 per cent CI for θ
F	4	2692.91	0.29	-0.16	-0.20, -0.12
G	4	2737.61	0.31	-0.11	-0.25, 0.03
H	4	2740.23	0.30	-0.12	-0.25, 0.02
I	4	2750.74	0.28	-0.19	-0.35, -0.03

F = k components mixture distribution of baseline effect and fixed treatment effect
G = k components mixture distribution of baseline effect and treatment effect
H = k components mixture distribution of baseline effect and treatment effect plus fixed randomization design
I = k components mixture distribution of baseline effect and treatment effect plus fixed randomization design and interaction of treatment and randomization design

Table 7.6 Estimates of effects for the variables of model I

Variable	Estimated log-relative risk	95 per cent CI
a) Randomization design	0.13 (0.04)	0.05, 0.21
b) Multiple interventions	-0.19 (0.08)	-0.35, -0.03
c) Interaction a and b	0.21 (0.05)	0.11, 0.31

The classification of trials to each component according to model I, which adjusts for randomization design and the interaction effect, is presented in table 7.7. There is some difference in specific log-relative risk for individual components ranging from -0.29 to -0.06. The 2nd and 4th components show similar figures of significant protective effect of the multiple interventions. Even the 1st and 3rd components show non-significant protective effects of the multiple interventions. The upper limits of the confidence intervals close to zero. There is also little variation between the four components with a standard deviation of 0.08.

The results obtained from this example show that there is evidence of slight protective effect of multiple interventions on smoking prevalence among the adults aged at least 40 years and having no clinical evidence of established cardiovascular disease. Baseline

heterogeneity is found. The NPML approach also presents evidence of different randomization designs providing the different effect of multiple interventions.

Table 7.7 NPML estimates of treatment effect distribution and classification of trials according to the 4 components of the maximum likelihood estimates of model I

K^{th} Component	Estimated log-relative risk (95 per cent CI)	Weight	Trial number
1	-0.06 (-0.20, 0.08)	0.143	2, 13
2	-0.29 (-0.37,-0.21)	0.218	1, 4, 10
3	-0.13 (-0.29, 0.03)	0.282	6, 7, 11, 14
4	-0.24 (-0.36, -0.12)	0.357	3, 5, 8, 9, 12

7.6 Summary

The GLMM shows that by assuming free-distribution for the random effects and with the procedure of NPML estimator via EM algorithm, information on heterogeneity between trials is easily provided from sources of treatment effect, some potential covariates and random effects. Estimated variance of random effects is obtained from the variation of baseline in the K component mixing distribution. Despite this benefit, care should be taken when interpreting treatment effect in terms of risk since some of the asymptotic generalizibility issues remain unsolved in the nonparametric approach.

Topic 8: Comparing GLM and GLMM approaches, application to meta-analyses involving cluster randomized trials

In the previous two topics, GLM and GLMM approaches applied for the meta-analyses related to cluster randomized trials are discussed individually. In this topic an evaluation is performed for these approaches in terms of methodology, heterogeneity information provided, model complexity and numerical results.

Section 8.1 compares different approaches in several aspects of their methodology. Section 8.2 provides a discussion on the heterogeneity information obtained from individual approaches and model complexity. Section 8.3 discusses strengths and limitations of the individual approaches. Section 8.4 discusses comparison of the approaches in numerical results. Finally, section 8.5 concludes and proposes the approach for quantitative synthesis of the binary outcome in meta-analyses involving cluster randomized trials.

8.1 Methodology comparison

In terms of methodology for different approaches, the issues to be considered are estimation of parameters and assumption requirements for the estimation, and computation procedure. Table 8.1 shows procedures and required assumptions to estimate overall treatment effect for different approaches in the meta-analysis involving cluster randomized trials with binary outcome. Three aspects of the treatment effect are considered: an estimated overall treatment effect ($\hat{\theta}$), standard error of $\hat{\theta}$ and a confidence interval for θ.

For the GLM, restricted maximum likelihood (REML) estimator is a common estimator suggested to estimate the overall log-relative risk of treatment effect. The estimator requires the assumption of a normal distribution of observed treatment effect that is conditional on the linear mixed model of covariate and random effects. The random effect distribution here is approximately normal. The GLMM use nonparametric maximum likelihood approach to estimate the overall log-relative risk of treatment effect. The observed number of response variable has a poisson distribution given covariates and random effects. The random effects distribution is left unspecified.

In GLM, the REML estimator also provides the standard error of $\hat{\theta}$ under the additional required assumption as for the estimation of overall treatment effect. But since the standard error calculation is based on asymptotic properties, it may be unreliable. Therefore, the standard error is not used to calculate a confidence interval for θ. For the GLMM, the NPML estimator also provides the standard error of $\hat{\theta}$ under the additional required assumptions as mentioned to estimate the overall treatment effect.

To calculate a confidence interval for θ, the standard method for a mean parameter requiring an asymptotic normal distribution of estimated overall treatment effect can be used for the GLMM. For the GLM, parametric bootstrapping estimation is used to produce the confidence interval. The lower and upper limits of the interval are obtained from smoothed percentiles of bootstrap distributions. The procedure does not require any assumption for producing the confidence interval.

Table 8.1 Procedure and required assumption for estimating overall treatment effect for different approaches in the meta-analyses of situation studied

Treatment effect	Approaches	
	GLM	GLMM
Overall effect (θ) • Estimation procedure • Required assumption	REML Normaldistribute observed treatment effect given covariate and random effects. Normal distribution of random effects.	NPMLE Normal/Poisson/Binomial distributed observed response given covariate and random effects. Unspecified distribution of random effects.
Standard error of estimated θ • Estimation Procedure • Required assumption	REML Independent estimated θ Normal distributed observed treatment effect given covariate and random effects. Normal distribution of random effects	NPMLE Independent estimated θ Normal/Poisson/Binomial distributed observed response given covariate and random effects. Unspecified distribution of random effects
Confidence interval for θ • Procedure • Required assumption	Parametric bootstrapping estimation No requirement	Standard method for mean parameter Asymptotic normal distribution of estimated treatment effect

The comparative information presented in table 8.1 is also applied to the estimation of covariates effects.

For the random effects, three measures are considered: an estimated variance $\hat{\tau}^2$, standard error of $\hat{\tau}^2$ and a confidence interval for τ^2. Procedures and assumption for estimating variance components for different approaches are presented in table 8.2. The GLM also employs the REML estimator to provide the estimated variance $\hat{\tau}^2$ and a standard error of $\hat{\tau}^2$ under the assumption of normal distributed random effects. It is not advisable to use the standard error to calculate a confidence interval for τ^2 because the calculation of standard error is performed under asymptotic properties, and thus the standard error of $\hat{\tau}^2$ may be unreliable. The parametric bootstrapping estimation is used instead to produce the confidence interval for τ^2 without any assumption requirement.

For the GLMM, a weighted variance method is used to estimate $\hat{\tau}^2$ without any assumption requirement. In fact the approach can provide a standard error of $\hat{\tau}^2$ and a confidence interval for τ^2 by bootstrapping estimation without requirement of any assumption. However, the software used for the approach in this study does not have programmes for the bootstrapping estimation to provide these figures.

Table 8.2 Procedures and required assumptions for estimating variance of random effects for different approaches in meta-analyses of situation studied

Random effect	Approach	
	GLM	GLMM
Variance (τ^2) • Estimation procedure • Required assumption	REML Normal distribution of random effects	Weighted variance method No requirement
Standard error of estimated τ^2 • Estimation procedure • Required assumption	REML Normal distribution of random effects	Bootstrapping estimation* No requirement
Confidence interval for τ^2 • Procedure • Required assumption	Smoothed percentiles of bootstrapping distribution No requirement	Smoothed percentiles of bootstrapping distribution* No requirement

* could be obtained by the approach but not available in the software used here

Table 8.3 shows the computation procedure, software available and the software used for different approaches. They all use the iterative procedure for computing. For the GLM, the RIGLS algorithm is used to implement the estimations as provided in the MLwiN software used in this study. However, as discussed(117), it produces results similar to those from other algorithms.

For the GLMM, by using the macro programme of 2- level poisson regression of S-plus software, the iterative EM-algorithm is the procedure provided in the programme. This algorithm has been proposed (124) to estimate maximum likelihood of the GLMM parameters for many years. It is commonly used in this area.

Several software are available for the two approaches. The STATA software does not provide complete results in the available commands. Meta-analysts need to write more programmes for obtaining the complete results of the approaches. Thus, different software are selected to analyze the data of individual approaches for this study. The software is selected under the criteria that it provides most of the results needed. In addition, a familiar and friendly software is preferably chosen. The choice for each approach is believed to be appropriate.

Table 8.3 Computation procedure and software available for different approaches

Computation	Approach	
	GLM	GLMM
• Procedure	Iterative (RIGLS algorithm for this study)	Iterative (EM-algorithm for this study)
• Software available	Meta Graphs, GLIM MLwiN* SAS Macro Suite STATA Macros	GLIM GLIMMIX STATA Macros, S-plus*

* software used in this study

8.2 Heterogeneity information for different approaches

Here the heterogeneity information obtained from each approach is discussed. The summaries are presented in table 8.4. The GLM approach produces the estimated overall treatment effect and confidence interval for θ. Since the observed treatment effect is conditional upon a linear mixed model of covariates and random effects, the approach can provide covariate effect of continuous and categorical variables at trial level and individual level with the confidence intervals for the true parameters. For the random effects information, the GLM also provides an estimated variance $\hat{\tau}^2$ and a confidence interval for τ^2 for the whole meta-analyses and subgroups of some categorical covariate variables. An example is the randomization design variable for the meta-analysis of mammographic screening trials with two categories. Here, for each category a variance of random effects can be specified.

The GLMM provides the most general results compared to the GLM approach. The results not provided here by the GLMM are confidence intervals for τ^2 and estimated variances of covariate subgroups. These deficiencies are due to incomplete results provided by the software rather than the approach.

In terms of interpretation and generalizability, the results obtained from the GLM performed under normality assumption are straightforward to interpret. The GLMM obtain the estimated treatment effects from a discrete mixing distribution. For ones who believe in smoothing distribution, this may make it difficult to interpret and make an inference on the results. However, the issue of misspecified and unproved normal distribution of random effects is still questionable, especially for the common case of meta-analysis with small number of trials.