Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Na Bo Date

Random Estimating Functions to Accommodate Heterogeneity in Meta-Analysis

By

Na Bo
Master of Science in Public Health

Biostatistics and Bioinformatics

John Hanfelt
(Thesis Advisor)

Rebecca Zhang
(Reader)

Random Estimating Functions to Accommodate Heterogeneity in Meta-Analysis

By

Na Bo

Bachelor of Economics
Tianjin University of Finance and Economics
2015

Thesis Committee Chair: John Hanfelt, Ph.D.
Reader: Rebecca Zhang, M.S.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2017

Abstract

Random Estimating Functions to Accommodate Heterogeneity in Meta-Analysis

By Na Bo

Introduction: Meta-analysis is defined here as the statistical analysis of a collection of analytic results for the purpose of integrating the findings (DerSimonian & Laird, 1986). A major concern in meta-analysis is heterogeneity among the studies contributing analytic findings. Failure to account for heterogeneity could lead to misleading conclusions in a meta-analysis. The aim is to use statistical approaches to derive a common estimated odds ratio that represents the common truth behind multiple similar studies.

Methods: To accommodate heterogeneity, we propose to add a random perturbation to each component estimating function. The advantages of this proposal over a random-effects model are that, under reparametrization, the random estimating function remains unbiased, remains subject to an additive perturbation, and has a variance that is well-governed and easy to evaluate.

Results: Our new method can capture between- table heterogeneity and produces a valid estimate of the log odds ratio. An advantage of our new method is that it can be applied to further meta-analysis studies under reparametrization, by simply applying the Delta Method.

Discussion: A major advantage of our random estimating equation method over existing random-effects methods is that our new method can be implemented into meta-analyses for any 1-1 transformation of the odds ratio. Unlike a random-effects model, however, our approach does not easily suggest a data generation mechanism, which makes it challenging to conduct a simulation study. The ways of generating random observations under our model of a randomly perturbed Mantel-Haenszel estimating function need to be explored further.

Random Estimating Functions to Accommodate Heterogeneity in Meta-Analysis

By Na Bo

Bachelor of Economics
Tianjin University of Finance and Economics
2015

Thesis Committee Chair: John Hanfelt, Ph.D.
Reader: Rebecca Zhang, M.S.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2017

CHAPTER 1: INTRODUCTION

1.1 Background

Meta-analysis is defined here as the statistical analysis of a collection of analytic results for the purpose of integrating the findings (DerSimonian & Laird, 1986). A major concern in meta-analysis is heterogeneity among the studies contributing analytic findings. Failure to account for heterogeneity could lead to misleading conclusions in a meta-analysis (Kontopantelis et al., 2013). To explore these ideas, suppose the meta-analysis consists of a collection of case-control studies, a series of 2×2 tables. The aim that is to use statistical approaches to derive a common estimated odds ratio that represents the common truth behind multiple similar studies. Complicating the statistical analysis, however, is the need to adjust for possible heterogeneity across the 2×2 tables. The two most widely used methods are the fixed-effect model (Mantel & Haenszel,1959) and the random-effect model (DerSimonian & Laird, 1986). The latter method adequately account for heterogeneity. However, problems arise under reparametrization, which we describe below.

1.2 Problem Statement

To accommodate heterogeneity in estimating a common log odds ratio in a meta-analysis, a popular approach is to use a random-effect model. Under this approach, each study has its own log odds ratio parameter, θ_i, with an additive random perturbation (Bhaumik et al., 2012; DerSimonian & Laird, 1986)

$$\theta_i = \theta + \epsilon_i,\ \epsilon_i\ iid\ F[0, \tau^2],$$

where $F[0, \tau^2]$ is a distribution with mean 0 and variance τ^2. Under this model, each study's log odds ratio is unbiased, in the sense that $E(\theta_i) = \theta$, with variance given by $var(\theta_i) = \tau^2$.

The above model is sensitive to the choice of parametrization and is limited to additive random perturbation. For example, when taking the square root of a study-specific log odds ratio, the reparametrized log odds ratio will be biased, $\mathrm{E}(\sqrt{\theta_i}) \neq \theta$. Moreover, the exponent of θ_i is subject to multiplicative, rather than additive perturbation:

$$e^{\theta_i} = e^{\theta} e^{\epsilon_i}.$$

Since random-effect models have properties that change drastically under reparametrization, we propose an alternative approach in the common situation where we are uncertain as to the singular choice of parameter which is subject to the additive random perturbations ϵ_i.

In this paper, we put forward a new approach, in which a random perturbation is added into each component estimating function. By applying the random perturbations to the component estimating functions rather than to the exposure effects directly, we achieve inference in meta-analysis that is invariant under reparametrization. Moreover, our approach reduces to the simple Mantel-Haenszel estimator when there is no heterogeneity among the 2×2 tables.

1.3 Purpose Statement

The purpose of this paper is to generate a random-perturbed estimating function to estimate the common log odds ratio in meta-analysis, which accommodates heterogeneity across studies. Under 1-1 reparametrization, the new random estimating function remains

unbiased, subject to an additive perturbation, and has a variance that is well-governed and easily evaluated under reparametrization. Estimators of the common log odds ratio, between-table variance and variance of the common log odds ratio are examined under two cases: independent binomials with large row totals but a small number of strata (i.e., non-sparse data); and independent binomials with a large number of strata but small row totals (i.e., sparse data). Analyses of real data are conducted to compare our new approach with the Mantel-Haenszel method and the DerSimonian & Laird method.

1.4 Dataset Description

Eight randomized clinical trials with an exercise or diet intervention of at least 6 months of duration among participants at high risk of diabetes are included as the dataset that we use in this paper (Orozco et al., 2008). Since the specific interventions, eligibility criteria and the length of follow-up ranging from 1 to 6 years varied among studies, heterogeneity was likely to be considerable.

We form 2 x 2 tables (a, b, c, d) using the outcomes data in Appendix 5 of Orozco et al. (2008). For example, in the Bo 2007 study, the incidence of diabetes was 3/169 in the exercise & diet intervention group and 12/166 in the control group, and so (a, b, c, d) entries are (3, 166, 12, 154).

Here are the 2 x 2 tables (a, b, c, d) data:

Bo 2007: 3, 166, 12, 154

Da Qing 1997: 58, 68, 89, 44

DPP 2002: 155, 924, 313, 769

DPS 2002: 44, 197, 72, 167

IDPP 2006: 47, 73, 73, 60

Kosaka 2005: 3, 99, 33, 323

Oldroyd 2005: 7, 30, 8, 24

Wing 1998: 5, 27, 2, 29

CHAPTER 2: LITERATURE REVIEW

2.1 Meta-analysis

The use of meta-analysis for research synthesis has become popular in medical research, where information on efficacy of a treatment is available from a number of clinical studies with similar treatment protocols. Meta-analysis is defined as the statistical analysis of a collection of analytic results for the purpose of integrating the findings. If we were to consider each individual study separately, then any one of the studies is either too small or too limited in scope to come to unequivocable or generalizable conclusions about the treatment effect. Combining findings across studies becomes an attractive way to show evidence of treatment efficacy. Meta-analysis in medical research often focuses on the odds ratio (Engels et al. 2000; Deeks 2002) between treatment and control groups with a binary indicator of efficacy (Bhaumik et al., 2012; DerSimonian & Laird, 1986).

2.2 Mantel-Haenszel Estimate(MH)

The Mantel-Haenszel estimator is popular, but considers no heterogeneity across studies. Let $y_i = (a_i, b_i, c_i, d_i)$ be the entries in a 2×2 contingency table. Unconditionally, the (a_i, b_i, c_i, d_i) are two independent bonmial distributions; if we were to condition on the row totals m_{i1}, m_{i2} and column totals s_{i1}, s_{i2}, then the (a_i, b_i, c_i, d_i) follow a non-central hypergeometric distribution with odds ratio $\psi = e^{\theta}$,

$$P(a_i = x | m_{i1}, m_{i2}; \psi) = \frac{\binom{m_{i1}}{x}\binom{m_{i2}}{s_{i1}-x}\psi^x}{P_0(\psi)},$$

$$\text{where } P_0(\psi) = \sum_{j=a}^{b} \binom{m_{i1}}{j}\binom{m_{i2}}{s_{i1}-j}\psi^j,$$

where x runs from $a = \max(0, m_{i2} - s_{i1})$ to $b = \min(m_{i1}, s_{i1})$.

It is easily shown that

$$E(a_i d_i) = \psi E(b_i c_i),$$

so that $g_i(y_i, \theta) = a_i d_i - e^{\theta} b_i c_i$ is a zero-mean estimating function for the log odds ratio, θ.

Suppose that there are k 2×2 contingency tables with a common odds ratio, and that a pooled estimating function is desired. The pooled Mantel-Haenszel estimating function is given by (Breslow & Liang, 1982),

$$g(y, \theta) = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} g_i(y_i, \theta) = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} (a_i d_i - e^{\theta} b_i c_i)/N_i,$$

$$\text{where } N_i = a_i + b_i + c_i + d_i.$$

The Mantel-Haenszel estimating function is unbiased, in the sense that $E\{g(y, \theta)\} = 0$, with variance that can be easily estimated by $var\{g(y, \theta)\} = k^{-1} \sum_{i=1}^{k} var\{g_i(y_i, \theta)\}$. By solving the equation $g(y, \theta) = 0$, the Mantel-Haenszel estimator of the log odds ratio is given by $\hat{\theta} = \log(\sum a_i d_i / N_i) - \log(\sum b_i c_i / N_i)$, which is consistent and asymptotically normal under mild regularity conditions (Liang, 1985).

The Mantel-Haenszel approach is popular because of its simplicity and wide application. The Mantel-Haenszel estimator is valid in both the large-strata setting (large N_i, small k) and the sparse data setting (small N_i, large k) (Robins, Breslow & Greenland, 1986); moreover, the marginal row and column totals $M_i = (m_{i1}, m_{i2}, s_{i1}, s_{i2})$ can be considered either fixed or random (McCullagh, 1991), and extra-binomial variation is allowed within tables (Liang, 1985). The chief drawback to the fixed-effects Mantel-Haenszel approach is its failure to take into account of the heterogeneity in a meta-analysis.

The specific form of the asymptotic variance of $\hat{\theta}$ depends on the above scenarios (Breslow & Liang, 1982; Liang, 1985; Robins, Breslow & Greenland, 1986; McCullagh,

1991). For example, in the sparse data setting, the asymptotic variance of $\hat{\theta}$ is given by (Breslow & Liang, 1982)

$$\text{avar}(\hat{\theta}) = \frac{\sum var(g_i)}{\left\{\sum E\left(-\frac{\partial g_i}{\partial \theta}\right)\right\}^2} = \frac{\sum g_i^2(y_i,\hat{\theta})}{\left(\sum \frac{e^{\hat{\theta}} b_i c_i}{N_i}\right)^2} = \frac{\sum g_i^2(y_i,\hat{\theta})}{\left(\sum \frac{a_i d_i}{N_i}\right)^2}.$$

In this paper, we used combined estimator to estimate the variance of the MH log odds ratio, which shows consistency in both the large-strata setting and the sparse data setting (Robins, Breslow & Greenland, 1986). The empirical MH variance estimator of log odds ratio is defined by

$$\text{kVar}^{\text{A}}(ln\psi_{MH}) = kV_E(\theta) = k\psi_{MH}^{-2} Var_E(\psi_{MH})$$

$$\text{where } \text{Var}_{\text{E}}(\psi_{MH}) = \frac{\sum_i (R_i - \psi_{MH} S_i)^2 / k}{\sum_i \frac{S_i}{i}},$$

$$\text{R}_{\text{i}} = \frac{a_i(m_{i2} - c_i)}{N_i},$$

$$\text{S}_{\text{i}} = \frac{c_i(m_{i1} - a_i)}{N_i},$$

$$\text{N}_{\text{i}} = m_{i1} + m_{i2}.$$

The symmetric version of Hauck estimator is given by (Breslow & Liang, 1982)

$$\text{V}_{\text{HS}} = \left[\frac{\left(\sum_i \frac{S_i^2}{w_i}\right)\left(\sum_k \frac{R_i^2}{w_i}\right)}{(\sum_i S_i)^2 (\sum_i R_i)^2}\right]^{\frac{1}{2}}$$

$$\text{where } \text{w}_{\text{i}} = \left(\frac{1}{a_i + 0.5} + \frac{1}{m_{i1} - a_i + 0.5} + \frac{1}{m_{i2} - c_i + 0.5} + \frac{1}{c_i + 0.5}\right)^{-1}.$$

The combined estimator is defined as

$$\text{V}_{\text{BL}} = \frac{N_+ V_{HS} + k^2 V_E}{N_+ + k^2}$$

$$\text{where } \text{N}_+ = \sum_i N_i.$$

2.3 DerSimonian & Laird Estimate (DSL)

The DerSimonian & Laird Method addresses two issues: (1) the assignment of weights that reflect the relative "value" of the information provided in a study in purpose of combining studies; (2) one may be using incommensurable studies to answer the same question (DerSimonian & Laird, 1986).

The DerSimonian & Laird approach assumes that there is a distribution of treatment effects and utilize the observed effects from individual studies to estimate this distribution. The approach allows for treatment effects to vary across studies and provides an objective method for weighting that can be made progressively more general by incorporating study characteristics into the analysis (DerSimonian & Laird, 1986).

Suppose that there is a series of k comparative clinical consisting of treatment and control groups in each study. The number of patients with the event in each group are independent binomial random variables with probabilities p_{i1} and p_{i2}, i indexing from 1 to k. The method divides the observed treatment effect in each study into two additive parts: the true treatment effect, θ_i, and the sampling error, e_i. The variance of sampling error is usually estimated by s_i^2 from the ith observed data. The true treatment effect θ_i in each study is assumed to be associated with the mean effect for a population of possible treatment evaluations, θ, and the deviation of the ith study's effect from the population mean, ϵ_i,

$$\theta_i = \theta + \epsilon_i, \text{ with } \epsilon_i \; iid \; F[0, \tau^2].$$

In estimating pooled log odds ratio, DSL implicitly assume that the row totals m_{i1}, m_{i2} are large. The table-specific empirical odds ratios can be calculated by

$$\mathrm{y_i} = \log\left(\frac{a_i d_i}{b_i c_i}\right)$$

with asymptotic variances well-approximated by

$$\mathrm{w_i^{-1}} = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}.$$

The variance of the random effects is estimated by

$$\tau^2 = \max\left\{0, \frac{\sum w_i(y_i - \bar{y}_w)^2 - (k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}\right\}$$

$$\text{where } \bar{\mathrm{y}}_w = \frac{\sum w_i y_i}{\sum w_i}.$$

It is easily to be shown that the pooled log odds ratio is estimated by

$$\hat{\theta}_{DSL} = \sum w_i^*(\tau^2) y_i / \sum w_i^*(\tau^2)$$

$$\text{with avar}\left(\hat{\theta}_{DSL}\right) \doteq \frac{1}{\sum \mathrm{w}_i^*(\tau^2)}$$

$$\text{where } \mathrm{w}_i^*(\tau^2) = (\tau^2 + w_i^{-1})^{-1}.$$

The DSL approach is popular since it considers heterogeneity across study and the pooled treatment effect and between-study variance can be estimated intuitively.

While attractive, the above model is sensitive to the choice of parametrization. Non-affine transformations of θ_i generally are biased, e.g.,

$$E(\theta_i^2) = \theta^2 + \tau^2 \neq \theta^2;$$

moments that can be difficult to evaluate without further knowledge of the perturbation distribution F, e.g.,

$$E\left(\sqrt{\theta_i}\right) = E\left\{(\theta + \epsilon_i)^{\frac{1}{2}}\right\}, var\left(\sqrt{\theta_i}\right) = var\left\{(\theta + \epsilon_i)^{\frac{1}{2}}\right\};$$

moreover, exponential of θ_i is subject to multiplicative, rather than additive perturbation, e.g.,

$$e^{\theta_i} = e^{\theta} e^{\epsilon_i}.$$

CHAPTER 3: METHODOLOGY

3.1 Study Design

To accommodate heterogeneity, we propose to add a random perturbation to each component estimating function:

$$\mathrm{g}_{\epsilon}(y,\theta) = \frac{1}{\sqrt{k}}\sum_{i=1}^{k} g_{\epsilon i}(y_i,\theta) = \frac{1}{\sqrt{k}}\sum_{i=1}^{k}\{g_i(y_i,\theta)+\epsilon_i\}, \epsilon_{\mathrm{i}}\ iid\ F[0,\tau^2],$$

where the perturbations ϵ_{i} and observations y are independent. It follows that the random estimating function above is unbiased, in the sense that $\mathrm{E}\{\mathrm{g}_{\epsilon}(y,\theta)\} = 0$, with variance larger than under the unperturbed model, $\mathrm{var}\{\mathrm{g}_{\epsilon}(y,\theta)\} = var\{g(y,\theta)\} + \tau^2$. If the ϵ_{i} were observed, then one could estimate θ by directly solving the equation $\mathrm{g}_{\epsilon}(y,\theta) = 0$, yielding a solution $\hat{\theta}_{\epsilon}$ with larger asymptotic variance than the naïve estimator $\hat{\theta}$ that ignores heterogeneity; for example, in sparse data setting,

$$\mathrm{avar}\left(\hat{\theta}_{\epsilon}\right) = \frac{k^{-1}\sum var(g_i) + \tau^2}{\left\{k^{-\frac{1}{2}}\sum E\left(-\frac{\partial g_i}{\partial\theta}\right)\right\}^2} > avar\left(\hat{\theta}\right).$$

The advantages of this proposal over a random-effects model are that, under reparametrization, the random estimating function proposed above remains unbiased, remains subject to an additive perturbation, and has a variance that is well-governed and easy to evaluate. To see this, consider a 1-1 transformation $\theta \longmapsto \eta(\theta)$, such that the inverse function $\theta(\eta)$ is differentiable. Mimicking what would occur to the fully parametric score function under this transformation, we define the random estimating function under the η parametrization as

$$\mathrm{g}^{*}_{\epsilon}(y,\eta) = \frac{\partial\theta(\eta)}{\partial\eta} g_{\epsilon}\{y,\theta(\eta)\},$$

where we see that $\mathrm{E}\{g_\epsilon^*(y,\eta)\} = 0$ and $\mathrm{var}\{g_\epsilon^*(y,\eta)\} = \left\{\frac{\partial\theta(\eta)}{\partial\eta}\right\}^2 var[g_\epsilon\{y,\theta(\eta)\}]$. Moreover, if the ϵ_i were observed, then the resulting estimator $\hat{\eta}_\epsilon$ of η, found by solving $g_\epsilon^*(y,\eta) = 0$, is invariant in the sense that $\hat{\eta}_\epsilon = \eta(\hat{\theta}_\epsilon)$, with asymptotic variance given by $\mathrm{avar}(\hat{\eta}_\epsilon) = \left\{\frac{\partial\eta(\theta)}{\partial\theta}\right\}^2 avar(\hat{\theta}_\epsilon)$, in accordance with the Delta Theorem.

3.2 Log Odds Ratio Estimation and Between Table Variance Estimation

3.2.1 Estimation Process of θ and τ^2

We can consider at least three cases: Case 1--Margins M_i of 2×2 tables are fixed; Case 2—independent binomials with large row totals m_{i1}, m_{i2}; Case 3—independent binomial with small row totals m_{i1}, m_{i2}. Case 2 and Case 3 will be evaluated in this paper.

To estimate θ and τ^2, we adopt the simple approach by Paule & Mandel (1982) to our random estimating function context. After initializing $\tau^2 = 0$, we iterate between the following two steps until the estimates of θ and τ^2 converge:

(a) Keeping the current estimate of τ^2 fixed, solve the following estimating equation for θ:

$$g_{w(\tau^2)}(y,\theta) = \frac{1}{\sqrt{k}}\sum_{i=1}^{k}\frac{w_i(\tau^2)\left(a_i d_i - e^\theta b_i c_i\right)}{N_i} = 0,$$

where $w_i(\tau^2)$ is a convenient choice of weight, discussed below, satisfying $w_i(0) = 1$. It follows that

$\hat{\theta}(\tau^2) = \log\{\sum w_i(\tau^2)a_i d_i/N_i\} - \log\{\sum w_i(\tau^2)b_i c_i/N_i\}$, which reduces to the usual Mantel-Haenszel estimator when $\tau^2 = 0$.

(b) Keeping the current estimate of θ fixed, solve the following estimating equation for τ^2:

$$\mathrm{h}(\mathrm{y}, \tau^2, \theta) = \sum_{i=1}^{k} \frac{g_i^2(y_i,\theta)}{\tau^2 + var\{g_i(y_i,\theta)|\epsilon_i;\theta\}} - (k-1) = 0,$$

where $\mathrm{var}\{\mathrm{g_i}(y_i,\theta)|\epsilon_i;\theta\}$ is a model, discussed below, of the variance of the estimating function $\mathrm{g_i}(y_i,\theta) = (a_i d_i - e^{\theta} b_i c_i)/N_i$, keeping the perturbation ϵ_i fixed.

The specific choices of $\mathrm{w_i}(\tau^2)$ and $\mathrm{var}\{\mathrm{g_i}(y_i,\theta)|\epsilon_i;\theta\}$ for the use in steps (a) and (b) will depend on the underlying model assumptions. Case 2 and Case 3 are considered below. Both assume that extra-binomial variation is not present within the 2×2 tables.

3.2.2 Case 2: Independent binomials with large row totals m_{i1}, m_{i2}.

A convenient weight that is the optimal unconditional weight when $\theta = 0$ is given by

$$\mathrm{w_i}(\tau^2) = \frac{E(-\frac{\partial g_{\epsilon i}}{\partial \theta};\theta=0)}{var(g_{\epsilon i};\theta=0)} = \left\{1 + \frac{\tau^2 N_i}{m_{i1} m_{i2} p_{i2}(1-p_{i2})}\right\}^{-1},$$

where, when the row totals are large and $\theta = 0$, we can safely substitute the crude estimator $\tilde{p}_{i2} = (a_i + c_i)/N_i$ for the unknown nuisance parameter $\mathrm{p_{i2}}$. When the row totals are large, Breslow & Liang (1982) suggest using the following approximation to the variance for use in step (b):

$$\mathrm{var}(\mathrm{g_i}|\epsilon_i;\theta) \doteq e^{2\theta} \frac{b_i^2 c_i^2}{N_i^2} \left(\frac{1}{a_i+\frac{1}{2}} + \frac{1}{b_i+\frac{1}{2}} + \frac{1}{c_i+\frac{1}{2}} + \frac{1}{d_i+\frac{1}{2}}\right).$$

In the addendum to Breslow & Liang (1982), it pointed out that is variance approximation has a 'curious lack of symmetry' under interchange of the rows or columns in the 2×2 table. A symmetrized version is available by briefly considering the

rescaled estimating function $g_i^* = e^{-\frac{\theta}{2}}(a_i d_i - e^{\theta} b_i c_i)/N_i$, where interchange of the rows or columns in the 2×2 table would have the effect of merely changing the sign of g_i^*. Using the geometric mean of the two available large-sample variance approximation of g_i^*, it follows that

$$\mathrm{var}(g_i|\epsilon_i;\theta) = e^{\theta} var(g_i^*|\epsilon_i;\theta) \doteq e^{\theta} \frac{a_i b_i c_i d_i}{N_i^2}\left(\frac{1}{a_i+\frac{1}{2}} + \frac{1}{b_i+\frac{1}{2}} + \frac{1}{c_i+\frac{1}{2}} + \frac{1}{d_i+\frac{1}{2}}\right).$$

3.2.3 Case 3: Independent binomials with small row totals m_{i1}, m_{i2}.

When the row totals within the 2×2 tables are not large, it is advisable to 'borrow strength' to better estimate the nuisance parameter p_{i2} in the unconditional weight (4). A simple shrinkage estimator of p_{i2} such as would arise from an Empirical Bayes approach when the prior distribution is $p_{i2} \sim Beta(\alpha, \beta)$, when $\theta = 0$ is given by

$$p_{i2}^* = \frac{a_i + c_i + \alpha^*}{N_i + \alpha^* + \beta^*},$$

where $\alpha^* = p^*(1-\gamma^*)/\gamma^*$ and $\beta^* = (1-p^*)(1-\gamma^*)/\gamma^*$, with

$p^* = \sum(a_i + c_i)/\left(\sum N_i\right)$ and γ^* obtained by solving the equation

$$q\{y, p^*, \gamma^*\} = \sum_{i=1}^{k} \frac{(a_i + c_i - N_i p^*)^2}{N_i p^*(1-p^*)\{1+(N_i-1)\gamma^*\}} - (k-1) = 0.$$

When $\theta \neq 0$, the unconditional variance for use in step (b) is (McCullagh, 1991, p. 277)

$$var(g_i|\epsilon_i;\theta) = \frac{m_{i1} m_{i2} e^{\theta}}{N_i^2}\{m_{i1} p_{i1}(1-p_{i1}) + m_{i2} p_{i2}(1-p_{i2}) + (p_{i1}-p_{i2})^2\},$$

where we can estimate the parameters p_{i1} and p_{i2} by

$$p_{i1}^{\dagger} = \frac{\frac{p_{i2}^{\dagger}}{1-p_{i2}^{\dagger}} e^{\theta}}{1 + \frac{p_{i2}^{\dagger}}{1-p_{i2}^{\dagger}} e^{\theta}}, \; p_{i2}^{\dagger} = \frac{c_i + \alpha_2^{\dagger}}{m_{i2} + \alpha_2^{\dagger} + \beta_2^{\dagger}},$$

with $\alpha_2^{\dagger} = p_2^{\dagger}(1-\gamma_2^{\dagger})/\gamma_2^{\dagger}$, $\beta_2^{\dagger} = (1-p_2^{\dagger})(1-\gamma_2^{\dagger})/\gamma_2^{\dagger}$, $p_2^{\dagger} = \sum c_i / (\sum m_{i2})$, and $\gamma_2^{\dagger}$ obtained by solving the equation

$$q\left(y, p_2^{\dagger}, \gamma_2^{\dagger}\right) = \sum_{i=1}^{k} \frac{\left(c_i - m_{i2}p_2^{\dagger}\right)^2}{m_{i2}p_2^{\dagger}\left(1-p_2^{\dagger}\right)\{1+(m_{i2}-1)\gamma_2^{\dagger}\}} - (k-1) = 0.$$

Alternatively, symmetry-corrected estimates of p_{i1} and p_{i2} might be preferred, given by the geometric means $\left(p_{i1}^{\dagger}p_{i1}^{\dagger\dagger}\right)^{\frac{1}{2}}$ and $\left(p_{i2}^{\dagger}p_{i2}^{\dagger\dagger}\right)^{\frac{1}{2}}$, where

$$p_{i1}^{\dagger\dagger} = \frac{a_i+\alpha_1^{\dagger}}{m_{i1}+\alpha_1^{\dagger}+\beta_1^{\dagger}}, p_{i2}^{\dagger\dagger} = \frac{\frac{p_{i1}^{\dagger\dagger}}{1-p_{i1}^{\dagger\dagger}}e^{-\theta}}{1+\frac{p_{i1}^{\dagger\dagger}}{1-p_{i1}^{\dagger\dagger}}e^{-\theta}},$$

with $\alpha_1^{\dagger} = p_1^{\dagger}(1-\gamma_1^{\dagger})/\gamma_1^{\dagger}$, $\beta_1^{\dagger} = (1-p_1^{\dagger})(1-\gamma_1^{\dagger})/\gamma_1^{\dagger}$, $p_1^{\dagger} = \sum a_i / (\sum m_{i1})$, and $\gamma_1^{\dagger}$ obtained by solving the equation

$$q\left(y, p_1^{\dagger}, \gamma_1^{\dagger}\right) = \sum_{i=1}^{k} \frac{\left(a_i - m_{i1}p_1^{\dagger}\right)^2}{m_{i1}p_1^{\dagger}\left(1-p_1^{\dagger}\right)\{1+(m_{i2}-1)\gamma_1^{\dagger}\}} - (k-1) = 0.$$

3.3 Variance Estimation of Log Odds Ratio

3.3.1 Variance Estimator of Log Odds Ratio in Case 2

We use the fact that

$$e^{\widehat{\theta}} - e^{\theta} = \frac{\sum w_i(\tau^2)(a_i d_i - e^{\theta} b_i c_i)/N_i}{\sum w_i(\tau^2) b_i c_i / N_i},$$

to obtain

$$avar(e^{\widehat{\theta}}) \doteq \frac{\sum w_i^2(\tau^2)\{\tau^2 + e^{\theta}\frac{a_i b_i c_i d_i}{N_i^2}(\frac{1}{a_i+\frac{1}{2}}+\frac{1}{b_i+\frac{1}{2}}+\frac{1}{c_i+\frac{1}{2}}+\frac{1}{d_i+\frac{1}{2}})\}}{\left\{\sum \frac{w_i(\tau^2)b_i c_i}{N_i}\right\}^2}.$$

By using the Delta Theorem, it yields

$$\text{avar}\left(\widehat{\theta}\right) \doteq V_2 = \frac{\sum w_i^2(\tau^2)\{\tau^2 + e^{\theta}\frac{a_i b_i c_i d_i}{N_i^2}(\frac{1}{a_i+\frac{1}{2}}+\frac{1}{b_i+\frac{1}{2}}+\frac{1}{c_i+\frac{1}{2}}+\frac{1}{d_i+\frac{1}{2}})\}}{\left\{\sum \frac{w_i(\tau^2)a_i d_i}{N_i}\right\}^2}.$$

3.3.2 Variance Estimator of Log Odds Ratio in Case 3

We borrow the idea of weight average of Empirical estimator and model based estimator from Breslow & Liang (1982) to generate the variance estimator.

A model-based estimator of the variance is given by

$$\text{avar}(\hat{\theta}) \doteq V_{3M} = \frac{\sum w_i^2(\tau^2)[\tau^2+\frac{m_{i1}m_{i2}e^{\theta}\{m_{i1}p_{i1}(1-p_{i1})+m_{i2}p_{i2}(1-p_{i2})+(p_{i1}-p_{i2})^2\}}{N_i^2}]}{e^{2\theta}\left\{\sum\frac{w_i(\tau^2)m_{i1}m_{i2}(1-p_{i1})p_{i2}}{N_i}\right\}^2},$$

where we substitute $(\mathrm{p}_{i1}^{\dagger}p_{i1}^{\dagger\dagger})^{\frac{1}{2}}$ and $(\mathrm{p}_{i2}^{\dagger}p_{i2}^{\dagger\dagger})^{\frac{1}{2}}$ for unknown $\mathrm{p_{i1}}$ and $\mathrm{p_{i2}}$.

Alternatively, an empirical estimator of the variance is given by

$$\text{avar}(\hat{\theta}) \doteq V_{3E} = \frac{\sum\frac{w_i^2(\tau^2)(a_ib_i-e^{\theta}c_id_i)^2}{N_i^2}}{\left\{\sum\frac{w_i(\tau^2)a_id_i}{N_i}\right\}^2}.$$

Following the strategy of Breslow & Liang (1982), a combined estimator that is consistent under both asymptotic scenarios, Case 2 and Case 3, is the weighted average

$$\mathrm{V_{Cj}} = \frac{(\sum N_i)V_2+k^2V_{3j}}{(\sum N_i)+k^2}, (j \in \{M, E\}).$$

CHAPTER 4: RESULTS

The meta-analysis of exercise and diet interventions to prevent type 2 diabetes (Orozeo et al., 2008) included two studies with small row totals, and six studies with large row totals, and so conceivably, we could use either Case 2 method or our Case 3 method. After initializing $\tau^2 = 0$, we iterated between the two steps discussed in section 3.2.1 until the difference in parameter estimate between estimates of parameters θ and τ^2 converged. We set the convergence condition as the current iteration and last iteration less than or equal to 0.0001.

4.1 Case 2 Results: Independent Binomials with large row totals m_{i1}, m_{i2}.

Table1

	theta	se(theta)	tau.sq
MH	-0.815	0.083	0.000
DSL	-0.795	0.098	0.008
random estimating equation	-0.819	0.084	0.406

Under Case 2, we obtained a convergence in the third iteration (Table 2). As expected, the estimated log odds ratios were similar using the MH method, DSL method, and our proposed method, but the estimated heterogeneity across tables differed considerably, as did the estimated standard errors. The estimated between-table variance under MH method, DSL method and our method are 0.000, 0.008 and 0.406, respectively. Since our method measures the between-table variance as the estimating function scale and the DSl method measures the between-table variance on the exposure effect scale, the two estimates of τ^2 are not comparable. The estimated standard errors under the MH method, DSL method and our method were 0.08332074, 0.09759384 and 0.0842608, respectively. The standard error under our random estimating equation approach was only slightly

larger than the standard error under the MH method, suggesting that the heterogeneity across studies in the meta-analysis was not large.

4.2 Case 3 Results: Independent Binomials with small row totals m_{i1}, m_{i2}.

Table 2

	theta	se(theta)	tau.sq
MH	-0.815	0.083	0.000
DSL	-0.795	0.098	0.008
random estimating equation	-0.818	0.084	0.267

Under Case 3, we obtained convergence in the third iteration (Table 2). The DSL method was intended for Case 2 only, and makes no adjustment for small row totals, and so the DSL results remained the same as in Case 2. The estimated between-table variance under MH method, DSL method and our method are 0.000, 0.008 and 0.267 respectively. The estimated $\tau^{2\prime}s$ show the same general conclusion as case 2. Compared with case 2, the estimated τ^2 under our method was smaller here.

The estimated log odds ratios under MH method, DSL method and our method were quit similar. The estimated standard errors under MH method, DSL method and our method (using our model-based method) were 0.083, 0.098 and 0.084, respectively. The standard error under our random estimating equation was only slightly larger than the standard error under the MH method.

From these results above, we conclude that our new method can capture between- table variance and obtain a valid estimates of the log odds ratio. An advantage of our new

method is that it can be applied to further meta-analysis studies under reparametrization by simply applying the Delta Method.

CHAPTER 5: Discussion

5.1 Strengths

The purpose of this report is to examine our new method of estimating the pooled log odds ratio in a meta-analysis, which unlike existing random-effects methods is not sensitive to the choice of parametrization. We considered two different cases: Case 2 consisted of independent binomials with large row totals, and Case 3 consisted of independent binomials with small row totals. From the results in Chapter 4, we see that our proposed method successfully captures the heterogeneity across studies and yields valid log odds ratio estimates under these two scenarios. As expected, the standard error of the estimated log odds ratio was larger under our random estimating equation method than the one obtained by the MH method, which naively ignores heterogeneity. A major advantage of our random estimating equation method over existing random-effects methods is our the new method can be implemented into meta-analyses under any 1-1 transformation $\theta \mapsto \eta(\theta)$, such that $\theta(\eta)$ is differentiable, by using the transformed estimating equation

$$g_{\epsilon}^{*}(y,\eta) = \frac{\partial\theta(\eta)}{\partial(\eta)} g_{\epsilon}\{y,\theta(\eta)\},$$

where $E\{g_{\epsilon}^{*}(y,\eta)\} = 0$ with $var\{g_{\epsilon}^{*}(y,\eta)\} = \left\{\frac{\partial\theta(\eta)}{\partial\eta}\right\}^{2} var[g_{\epsilon}\{y,\theta(\eta)\}]$. If ϵ_i were observed, then the resulting estimate of η_ϵ can be found by solving $g_{\epsilon}^{*}(y,\eta) = 0$, which is invariant. Moreover, the estimated asymptotic variance of $\hat{\eta}_\epsilon$ can be easily solved.

5.2 Limitations

While our proposed approach improved upon the MH method by incorporating possible heterogeneity across the 2×2 tables, the resulting standard error of the estimated log odds

ratio was only slightly larger under our method compared to the MH method. This may be because that the data set that we used here was not a purely Case 2 or Case 3 data set. We can see that the row totals in Oldroyd's 2005 study and Wing's 1998 study were relatively small compared with six other studies. This may affect the performance of our new method. In future research, we will examine this issue.

In future research, we also will conduct a simulation study. This will be challenging, since the proposed method is not primarily used as a data generation process. The ways of generating random observations (a,b,c,d) under our model of a randomly perturbed Mantel-Haenszel estimating function need to be explored further.

REFERENCES

Breslow, N. E. and Liang, K. Y. (1982). The Variance of the Mantel-Haenszel Estimator. *Biometrics* **38**, 943-952.

Cordeiro, G. M. & McCullagh, P. (1991). Bias correction in generalized linear models. *J.R. Statist. Soc.* **B 53**, 629-43.

Deeks, J. (2002). Issues in the Selection of a Summary Statistic in Meta-Analysis of Clinical Trials With Binary Outcomes. *Statistics in Medicine*, 21, 1575-1600.

Bhaumik, D. K., Amatya, A., Normand, S. T., Greenhouse, J., Kaizar, E., Neelon, B., and Gibbons, R.D. (2012). Meta-Analysis of Rare Binary Adverse Event Data. *Journal of the American Statistical Association*, 107:498, 555-567, DOI: 10.1080/01621459.2012,664484.

Engels, E., Schmid, C., Terrin, N., Olkin, I., and Lau, J. (2000). Heterogeneity and Statistical Significance in Meta-Analysis: An Empirical Study of 125 Meta-Analyses. *Statistics in Medicine*, 19, 1707-1728.

Robins, J., Breslow, N., Greenland, S.. Estimators of the Mantel-Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models. *Biometrics*, Vol. 42, No. 2 (Jun., 1986), pp. 311-323.

Liang, K.Y. (1985). Odds Ratio Inference With Dependent Data. *Biometrics*, Vol. 72, No. 3 (Dec., 1985), pp. 678-682.

Kontopantelis E., Springate DA., Reeves D. (2013). A Re-Analysis of the Cochrane Library Data: the Dangers of Unobserved Heterogeneity in Meta-analyses. *PLoS ONE*, 8, 1-12.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from

retrospective studies of disease. *Journal of the National Cancer Institute* 22, 719-748.

Orozco LJ., Buchleitner AM., Gimenez-Perez G., Roqué i Figuls M., Richter B., Mauricio D. (2008). Exercise or exercise and diet for preventing type 2 diabetes mellitus. *Cochrane Database of Systematic Reviews* 2008, Issue 3. Art. No.: CD003054, DOI: 10.1002/14651858.CD003054.pub3.

Paule, R., and Mandel, J. (1982). Consensus Values and Weighting Factors. *Journal of Research of National Bureau Standard*, 87, 377-385.

DerSimonian, R. and Laird, N. (1986). Meta-Analysis in Clinical Trials. *Controlled Clinical Trials*, 7:177-188(1986).