Hierarchical Random-Effect Meta-Analysis of Binary Events
Investigating the Relation between Treatment Effect and
Underlying Risk

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy in the Graduate School of The Ohio State
University

By

Jingjing Yan, B.S., M.S.

Graduate Program in Statistics

The Ohio State University

2014

Dissertation Committee:

Eloise E. Kaizar, Ph.D., Co-Advisor
Steven N. MacEachern, Ph.D., Co-Advisor
Catherine A. Calder, Ph.D.
Elizabeth A. Stasny, Ph.D.
© Copyright by
Jingjing Yan
2014
Abstract

Meta-analysis is a statistical method used to quantitatively synthesize findings from independent but related studies. It has become increasingly important in medical research.

Using common moment-based meta-analytical approaches for studies with binary events, such as the Mantel-Haenszel method or the DerSimonian-Laird method, one would first choose a summary metric (e.g., relative risk, odds ratio or risk difference) and then “combine” across studies. The choice among different summary metrics is not straightforward given their different mathematical properties, ease of interpretation and variability across studies. We propose a flexible measure that allows estimation of the measure that “best” fits the data.

The accuracy of the moment-based estimators has been questioned by many. Alternative approaches are the likelihood-based methods that model the binomial structure of the data. The difficulty with these approaches lies in the model specification. This dissertation examines current hierarchical models and proposes new methods for hierarchical model selection.

By comparing conditional models to joint models, we conclude that the conditional approach that eliminates some nuisance parameters is convenient, but it neither agrees with other conditioning procedures nor captures the structure of the data as a full joint model with meaningful parameters does.
We extend the full joint model with a more flexible modeling structure, which is constructed based on a flexible measure. Instead of forcing an a priori decision among the common summary measures or a pre-fixed model, the flexible structure allows the data to decide which measure or model is the “best” for combining the studies. We look for the best way of describing the probability distribution for studies across a two-dimensional space (probability of an event in the control group and probability of an event in the treatment group). Inference on this scale is straightforward to interpret and can be easily transformed to any scale of interest.

The flexible modeling framework permits us to acknowledge the uncertainty in the model specification, but the intervals under the classical setting fail to account for it. We extend the methods to a fully Bayesian analysis to account for the full uncertainties in model selections and parameter estimations.
For my Mom Xiaoying Zeng, my Dad Qingbo Yan, and my Grandma Fengzhi Liang.

Thank you for your love and endless support.
Acknowledgments

It is hard to believe that my Ph.D. journey is nearing its end. Writing this dissertation was probably the hardest thing I have ever done in my life. I would not be standing at this point without the help from many people.

First and foremost, I thank my advisors, Dr. Eloise Kaizar and Dr. Steven MacEachern. They have taught me so much in the past three years. I know I would never be able to do this without them. I probably won’t miss all the 10/12-hour work days in grad school, but I will miss our weekly meetings so much. And I thank my dissertation committee members Dr. Catherine Calder and Dr. Elizabeth Stasny, for their valuable opinions. This work is also supported by NSF grant DMS-1209194.

Thank you Mom, Dad, and Grandma, for your love and support. Latex doesn’t support Chinese, so I inserted a picture with Chinese for them.

爸爸、妈妈、姥姥， 我爱你们。谢谢你们一直以来的支持和鼓励。

It was my undergrad thesis advisor Dr. Chunjie Wu who encouraged me to come to the US for graduate school. Thank you for believing in me, Wu Laoshi.

When I decided to come to OSU for grad school, I did not have a clue that I would meet my future husband here. Grant Schneider, you are the best thing that
ever happened to me. Thank you, Judy, Bill, the Schneider and the Offenberger family for the love you’ve given me. I can’t wait to be part of the family “officially”.

I would also like to thank my fellow students and friends Sarah Anderson, Steve Bamattre, John Lewis, and everyone else who accompanied and helped me on this grand journey.
Vita

December 17, 1986 ........................ Born - Yantai, China

2009 ...................................... B.S. Statistics

2011 ...................................... M.S. Statistics

2011-present ....................... Graduate Teaching Associate,
The Ohio State University.

Fields of Study

Major Field: Statistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Overview</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Examples</td>
<td>6</td>
</tr>
<tr>
<td>1.2.1 Lamotrigine Add-on</td>
<td>6</td>
</tr>
<tr>
<td>1.2.2 Cisapride</td>
<td>7</td>
</tr>
<tr>
<td>1.2.3 Endoscopic Hemostasis</td>
<td>8</td>
</tr>
<tr>
<td>1.2.4 Aprotinin</td>
<td>9</td>
</tr>
<tr>
<td>1.3 Problems to be Addressed</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Organization of This Dissertation</td>
<td>11</td>
</tr>
<tr>
<td>2. Literature Review</td>
<td>13</td>
</tr>
<tr>
<td>2.1 Heterogeneity</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Choice of a Measure of the Effect</td>
<td>16</td>
</tr>
<tr>
<td>2.3 Two-Stage Methods</td>
<td>20</td>
</tr>
<tr>
<td>2.3.1 Fixed-Effect Methods</td>
<td>21</td>
</tr>
<tr>
<td>2.3.2 Random-Effect Methods</td>
<td>23</td>
</tr>
<tr>
<td>2.3.3 Numerical Problems for Two-Stage Methods with Rare Events</td>
<td>25</td>
</tr>
</tbody>
</table>
2.4 Likelihood-based Methods ........................................... 29
  2.4.1 Approximate Models ........................................... 29
  2.4.2 Full Models ..................................................... 30
2.5 Bayesian Methods .................................................. 35
2.6 Summary ............................................................. 37

3. A Full Conditioning Approach and Comparisons between Conditional and Joint Models .................................................. 39
  3.1 Models for an Individual Study .................................... 40
    3.1.1 A Joint Model .................................................. 41
    3.1.2 A Conditional Model ......................................... 42
  3.2 Models for Multiple Studies ....................................... 42
    3.2.1 Joint Random-Effect Models .................................. 42
    3.2.2 The Non-central Hypergeometric Conditional Model .......... 46
    3.2.3 A Full Conditional Model ..................................... 46
  3.3 Approximating the Likelihood Functions and Estimating the Parameters .................................................. 48
  3.4 Illustration and Comparisons ...................................... 51
    3.4.1 A Numerical Example .......................................... 51
    3.4.2 Endoscopic Hemostasis Data .................................. 53
  3.5 Conclusions .......................................................... 55

4. A Flexible Measure and Two-Dimensional Distribution for Combining Studies in Meta-Analysis of Binary Events ...................... 57
  4.1 A Flexible Measure of Effect ...................................... 57
    4.1.1 Approximate Distribution ..................................... 60
    4.1.2 Illustration ..................................................... 61
    4.1.3 Quantifying the Heterogeneity ................................ 64
  4.2 Flexible Two-Dimensional Inference ................................ 69
    4.2.1 Random-effect Hierarchical Model ............................ 69
    4.2.2 Implementation ............................................... 72
    4.2.3 Simulation Study .............................................. 75
  4.3 Prediction ............................................................ 77
  4.4 Examples ............................................................. 80
  4.5 Summary ............................................................. 87

5. A Bayesian Extension to the Flexible Two-Dimensional Inference .... 95
  5.1 A Fully Bayesian Hierarchical Model .............................. 95
    5.1.1 Hyper-Parameters .............................................. 96
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Studies on the efficacy of Lamotrigine add-on for the treatment of drug-resistant partial epilepsy (any dose, any age) (Ramaratnam et al., 2001)</td>
<td>7</td>
</tr>
<tr>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>Studies on the efficacy of Cisapride on the treatment of functional (nonulcer) dyspepsia (Makambi, 2004)</td>
<td>8</td>
</tr>
<tr>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>The observed $2 \times 2$ table for study $i$</td>
<td>21</td>
</tr>
<tr>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>An observed $2 \times 2$ table of study $i$ with continuity correction.</td>
<td>27</td>
</tr>
<tr>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>The maximum likelihood estimates of $\theta$ and $\tau^2$ and the 95% confidence intervals using the non-central hypergeometric (NHG) and full conditional (FCOND) models.</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>Common measures, transformations, and estimates for a meta-analysis of binary events.</td>
<td>59</td>
</tr>
<tr>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>Special cases of the new summary metric $\theta(\alpha)$.</td>
<td>60</td>
</tr>
<tr>
<td>4.3</td>
<td></td>
</tr>
<tr>
<td>Estimation results for the simulation studies. Logistic denotes the random-effect logistic regression model and GLMM denotes the random-effect GLMM on the scale of $\theta(\alpha)$. $\pi_t$, $\pi_c$ (estimated by the medians of the distributions of $\pi_{ti}$ and $\pi_{ci}$) are the comparable parameters of the two models.</td>
<td>76</td>
</tr>
<tr>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>Estimation results for the four examples. Logistic denotes the random-effect logistic regression model and GLMM denotes the random-effect GLMM on the scale of $\theta(\alpha)$. $\pi_t$, $\pi_c$ and $\tau^2$ are the comparable parameters of the two models.</td>
<td>83</td>
</tr>
</tbody>
</table>
4.5 Prediction results for the treatment group event rate $\pi_t$ given the control group event rate $\pi_c^*$, and the treatment group number of events $r_t$ given $n_t^*$. The known information $n_c^*$, $\pi_c^*$, and $n_t^*$ are from a chosen study in the original datasets.

5.1 Results from using the Metropolis-Hastings method for fitting a fully Bayesian meta-analysis using the flexible GLMM for the Aprotinin dataset. Prior 1 is the initial prior distributions; $\mu$ has a more information prior distribution in prior 2.

C.1 Evaluation of the heterogeneity in different summary metrics using Cochran’s $\chi^2$ test and the Q-statistic.

D.1 Empirical correlation coefficients of the observed log odds ratio and the reciprocal of its sampling variance (one study).
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Odds ratios and 95% confidence intervals for the Lamotrigine add-on studies</td>
<td>10</td>
</tr>
<tr>
<td>3.1</td>
<td>Grid evaluation of $L_1^{NCH}(\theta, \tau^2)$ and the corresponding time used at different grid widths.</td>
<td>50</td>
</tr>
<tr>
<td>3.2</td>
<td>Comparison of the non-central hypergeometric likelihood and the full conditional model likelihood functions at given parameter values.</td>
<td>52</td>
</tr>
<tr>
<td>3.3</td>
<td>The profile log likelihood functions of $\theta$ and $\tau^2$ for the endoscopic hemostasis dataset. The horizontal dashed lines are the maximum log likelihood minus 1.92, indicating the confidence intervals. The vertical dotted lines show the MLEs. The non-central hypergeometric method is denoted by NHG and the full conditional model by FCOND.</td>
<td>53</td>
</tr>
<tr>
<td>4.1</td>
<td>L’Abbé plots on the risk difference scale. Each study is represented by a circle, the size (area in 1/16 inches) of which is proportional to its precision (inverse variance of the risk difference). The diagonal line is an indication of no effect ($\pi_t = \pi_c$).</td>
<td>62</td>
</tr>
<tr>
<td>4.2</td>
<td>Confidence interval plots of the 11 Lamotrigine add-on trials on the $\theta(\alpha)$ scale for $\alpha = 0, 0.5$ and 1. The studies are ordered by increasing observed control group event rates.</td>
<td>65</td>
</tr>
<tr>
<td>4.3</td>
<td>Confidence interval plots of the 64 Aprotinin trials on the $\theta(\alpha)$ scale for $\alpha = 0, 0.5$ and 1. The studies are ordered by increasing observed control group event rates. A continuity correction of 0.5 is used when necessary in the standard error calculation.</td>
<td>66</td>
</tr>
</tbody>
</table>
4.4 $I^2$ and $I^2_{AI}$ calculated from the Lamotrigine add-on and the Aprotinin dataset. The vertical lines show the $\alpha$ values that result in the smallest heterogeneity measured by $I^2$ and $I^2_{AI}$. . . . . . . . . . . . . . . . . . 68

4.5 A contour plot of log odds ratio vs. log odds in the control group under the bivariate normal assumption. The true log odds in the control group is $-2.21$, and the true log odds ratio is $0.23$. Values of the other parameters are $\pi_t = 0.3$, $\pi_c = 0.1$, $\tau^2 = 1$, $\omega^2 = 0.5$, and $\rho = 0.9$. . . . . . 71

4.6 Contour plots of $\theta(\alpha)$ vs. $\mu(\alpha)$. The normality assumption is made for $\alpha = 0.7$. The true parameters are $\pi_t = 0.3$, $\pi_c = 0.1$, $\tau^2_\alpha = 1$, $\omega^2_\alpha = 0.5$, and $\rho_\alpha = 0.9$. The plots are produced through simulation and kernel density estimation. . . . . . . . . . . . . . . . . . . . . . 73

4.7 Profile log likelihood function of $\alpha$. The dotted line indicates the $95\%$ likelihood-based confidence intervals defined in (4.15). . . . . . . . . . . . . . . . 81

4.8 Estimated predictive distributions of $(\theta_i(\alpha), \mu_i(\alpha))$ at $\hat{\alpha}$ with predictions for $\theta_{new}(\hat{\alpha})$ at given $\mu^*(\hat{\alpha})$. . . . . . . . . . . . . . . . . . . . . . 84

4.9 Lamotrigine add-on. Predictions on $\pi_t$ for different values of $\pi_c$. The points are the observed data. The black solid line represents the predicted values and the dashed red lines show the $95\%$ prediction band. The diagonal line indicates no effect. $\pi_t > \pi_c$ is preferable. . . . . . . 88

4.10 Cisapride. Predictions on $\pi_t$ for different values of $\pi_c$. The points are the observed data. The black solid line represents the predicted values and the dashed red lines show the $95\%$ prediction band. The diagonal line indicates no effect. $\pi_t > \pi_c$ is preferable. . . . . . . . . . . . . . . 89

4.11 Endoscopic Hemostasis. Predictions on $\pi_t$ for different values of $\pi_c$. The points are the observed data. The black solid line represents the predicted values and the dashed red lines show the $95\%$ prediction band. The diagonal line indicates no effect. $\pi t < \pi c$ is preferable. . . . . . . 90

4.12 Aprotinin. Predictions on $\pi_t$ for different values of $\pi_c$. The points are the observed data. The black solid line represents the predicted values and the dashed red lines show the $95\%$ prediction band. The diagonal line indicates no effect. $\pi_t < \pi_c$ is preferable. . . . . . . . . . . . . . . 91

xv
4.13 Prediction for the number of events in the treatment arm \( (r_t) \) of a future study given the number of patients in both arms \( (n_t, n_c) \) and the control group risk \( (\pi_c) \). The blue and green points indicate two prediction intervals of \( r_t \) that are closest to the 95% confidence level.

5.1 Prediction curve of \( \pi_t \) given values of \( \pi_c \) using results from the fully Bayesian model (priors 1). The plots on the side and on the bottom show the posterior distributions of the medians of \( \pi_t \) and \( \pi_c \).

5.2 Prediction curve of \( \pi_t \) given values of \( \pi_c \) using results from the fully Bayesian model (priors 1) except that \( \alpha \) is fixed at 0.595. The plots on the side and on the bottom show the posterior distributions of the medians of \( \pi_t \) and \( \pi_c \).

D.1 The relationship between the true log odds ratio and the empirical correlation coefficient.

E.1 Mean effect estimates of the non-central hypergeometric model, the DL model and the logistic model. The purple dotted lines indicate the true effect \( \theta = 0 \) and \( \hat{\theta} = \theta \) in the two plots, respectively. Estimates closer to the purple line are more desirable.

E.2 Bias of the mean effect estimates of the non-central hypergeometric model, the DL model and the logistic model at \( \rho = 0 \) for \( \theta = (-3, -1, 0, 1, 3) \). The purple dotted line indicates no bias.

E.3 Between-study variance estimates of the non-central hypergeometric model, the DL model and the logistic model. The purple dotted line indicates the true variance \( \tau^2 = 1 \). Estimates closer to the purple line is more desirable.
Chapter 1: Introduction

1.1 Overview

It is unusual for a scientific or a policy question to be answered by a single study. Multiple experiments are carried out to investigate the same phenomenon for many reasons: researchers are not aware of what their colleagues are working on, they extend the previous experiments to a larger sample or a different group of subjects, or they are simply suspicious of the previous results. In social science or medical research, repetitive studies are carried out to compare the effects of different interventions. When confronted with evidence from multiple studies, which often express conflicting findings, it becomes inevitable for the decision-makers to aggregate evidence from a variety of sources.

A quantitative synthesis of multiple studies that are different but related is known as meta-analysis. It is commonly carried out as part of a systematic review or research synthesis, which is a procedure used to produce an up-to-date summary of existing evidence on a particular topic (Sutton et al., 2002). A systematic review should be considered as a research design procedure, which requires careful consideration of a study protocol involving not only the statistical analysis, but also the
problem formulation, data collection, and interpretation of results. This dissertation focuses on meta-analytical modeling and inference, but it should be a common understanding that even the best statistical models cannot save a poorly designed study. Guidelines for performing systematic reviews are carefully written (Egger et al., 2008; Hedges and Cooper, 1993; Higgins and Green, 2011) and have become widely accepted, which is evidenced by the development of the Cochrane Database of Systematic Reviews, a source of high-quality health care and health policy synthesized evidence (http://www.cochrane.org/). Most of the Cochrane reviews are based on randomized controlled and clinical controlled trials. The development of the Cochrane Collaborate has led to similar developments for the synthesis of non-randomized observational studies and in the social science context, e.g. the Campbell systematic reviews summarizing evidence in crime and justice, education, social welfare and so forth (http://www.campbellcollaboration.org/).

The effort of dealing quantitatively with multiple observations traces back to the French mathematician Blaise Pascal who developed mathematical methods for games of chance used for gambling in the 17th century (O’Rourke, 2007). In addition to deciding the possible outcomes in gambling, the approaches were also employed by the mathematicians and astronomers to look for the best way to compare and combine astronomical observations. For example, the method of least squares was developed by French statistician Legendre to combine data from different observatories where the errors were known to be different (Chalmers et al., 2002). Observations from a study and summarized results from other studies were not clearly defined and differentiated back then. The ideas of observations, errors in the observations and the combination
of the observations were addressed in the 18th and 19th centuries by scientists such as Gauss and Laplace (Airy, 1861).

Although there were examples of research synthesis dating back to the end of the 19th century, for example Herber Nichols review on psychology in 1891, it was not until the 20th century when the modern research synthesis that we know today started to develop (Chalmers et al., 2002). The earliest recognized example of meta-analysis for clinical trials is Karl Pearson’s 1904 study on typhoid vaccine effectiveness (Sutton and Higgins, 2008), in which he used a collection of 11 studies and addressed several statistical questions that are still of interest to this day, such as the collection of data, randomization, heterogeneity, choice of measures and guidance on a future study (Pearson, 1904; Shannon, 2008):

- Pearson expressed his concern on the self-selection into the vaccine group and recommended carrying out an experiment (the concept of randomization had not been formally addressed yet).

- The data Pearson had was a collection of $2 \times 2$ tables, and he chose the correlation between the vaccine inoculation and typhoid mortality from each study as a measure of effect.

- Pearson looked to explain why the correlation values from each study contained “extreme irregularity” (heterogeneity).

- Pearson was especially interested in using the current results to guide future research. For example, he recognized the need for a controlled trial and proposed to only inoculate every second volunteer.
During the first half of the 20th century, the combination of the results of different studies was addressed by many, including Fisher. In his famous book *Statistical Methods for Research Workers*, Fisher presented a technique for combining p-values from independent tests of the same hypothesis (Fisher, 1932). Applications were seen in the fields of education, physics and agriculture (Birge, 1932; Yates and Cochran, 1938). The criteria for a literature search, study selection, data collection and statistical analysis started to be addressed as well (Winkelstein, 1998). In the middle of 20th century, an explosion of published studies, especially in social science research, increased the need for methods to quantitatively synthesize results from different studies. Chalmers et al. (2002) listed some of the problems the researchers were dealing with: 345 studies on the effects of interpersonal expectations on behavior, 725 estimates of the relation between class size and academic achievement and so forth. Gene Glass first used the name “meta-analysis” in his 1976 paper, referring to “the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings” (Glass, 1976).

An important leap forward concerning the application of meta-analysis in medical research also happened in the 1970s. Peter Elwood, Archie Cochrane and their colleagues were working on the first randomized trial to determine whether aspirin is protective of patients after heart attack (myocardial infarction). The first trial showed a non-significant difference between the aspirin group and the placebo group, which led to five more trials. All six of the trials indicated that aspirin lowered the mortality rate in the aspirin group, but none of the effects was statistically significant. To synthesize the available evidence, they conducted a meta-analysis and reported a weighted overall effect that indicated a significant beneficial effect of aspirin (Elwood
et al., 1974; Elwood, 2006). Inspired by their work, Yusuf et al. (1984) went on and conducted further research on the usage of meta-analysis in medical research. The authors provided detailed examples and reviews explaining why systematic overviews of randomized trials are needed and addressed the problems involved (Yusuf et al., 1985).

Large-scale randomized evidence is needed in medical research because the true difference between the two treatments is often moderate. Such differences can only be detected when there are relatively small biases and/or random errors present, which requires randomized trials with a large number of patients. When such large trials are not available, evidence can be obtained from an overview of multiple available randomized trials. A medical assumption that makes such overview reasonable is: “if for some identifiable category of patients, a moderate difference between two treatments exists, then for other groups of patients, this difference could be smaller or bigger but it is unlikely to be reversed” (Peto et al., 1995). In other words, the treatment effect differences of various trials (e.g. trials from different countries, different populations, or different age groups) are more likely to be in the sizes of effects, rather than in the direction. This is the fundamental assumption underlying meta-analysis in medical decision making.

By combining studies, meta-analysis provides more power to detect a treatment effect compared to a single study, especially when the effect is moderate (Peto et al., 1995). When findings of different studies conflict, meta-analysis can be used to estimate a mean effect, to identify studies related to a beneficial (or harmful) effect, and to estimate the magnitude of the effect. Another advantage of meta-analysis lies in its ability to formally assess the generalizability of findings from one type of
study to another (Borenstein et al., 2009). Many studies, although targeting the same intervention, have different characteristics (such as patient population and study design) which may cause differences in the treatment effect of interest. These possible differences can be investigated by meta-analysis.

Combining observational studies is performed less frequently, but is beginning to draw researchers’ attention, as it can be used to evaluate an exposure-disease relationship in medical research and post-marketing effectiveness or side effects in drug development. Differences between observational studies in a meta-analysis are often larger than those between randomized controlled trials (RCTs), and there exists potential confounding in observational studies. Therefore, observational study meta-analysis is rarely designed to estimate an overall effect, but rather to investigate patterns and reasons for different results (Stroup et al., 2000).

1.2 Examples

We use four examples to illustrate the methods developed in this dissertation. They are all collections of randomized controlled studies with binary outcomes (zero/one) in medical research. The examples are chosen to illustrate a range of different characteristics such as the numbers of studies, the type of endpoints (beneficial or harmful) and the extent of heterogeneity.

1.2.1 Lamotrigine Add-on

In the first example, 11 randomized controlled studies reported on the efficacy of Lamotrigine add-on (any dose, adults or children) for the treatment of drug-resistant partial epilepsy. The results are summarized in Table 1.1. An event is defined as 50% or more reduction in seizure frequency. The Duchowny study recruited only children
Table 1.1: Studies on the efficacy of Lamotrigine add-on for the treatment of drug-resistant partial epilepsy (any dose, any age) (Ramaratnam et al., 2001)

<table>
<thead>
<tr>
<th>Study</th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. Events</td>
<td>No. Patients</td>
</tr>
<tr>
<td>Binnie 1989</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Jawad 1989</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Loiseau 1990</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Matsuo 1993</td>
<td>33</td>
<td>143</td>
</tr>
<tr>
<td>Schapel 1993</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Schmidt 1993</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Smith 1993</td>
<td>4</td>
<td>41</td>
</tr>
<tr>
<td>Messenheimer 1994</td>
<td>10</td>
<td>46</td>
</tr>
<tr>
<td>Boas 1996</td>
<td>6</td>
<td>30</td>
</tr>
<tr>
<td>Duchowny 1999</td>
<td>41</td>
<td>98</td>
</tr>
<tr>
<td>Naritoku 2007</td>
<td>49</td>
<td>121</td>
</tr>
</tbody>
</table>

(199 patients, 25% of the total number of patients in the 11 studies) and the other studies recruited only adults.

More events are preferred for these studies since the outcome is beneficial. The observed event rates are mostly small, although there are no zero-arms (an arm refers to a group of patients in a clinical trial) present.

1.2.2 Cisapride

Cisapride is a gastroprokinetic agent used in the treatment of functional (nonulcer) dyspepsia, which is characterized by a variety of upper abdominal symptoms in the absence of an organic disease. Thirteen randomized controlled trials are collected to study the efficacy of Cisapride (Makambi, 2004), whose results are shown in Table 1.2.
Table 1.2: Studies on the efficacy of Cisapride on the treatment of functional (nonulcer) dyspepsia (Makambi, 2004)

<table>
<thead>
<tr>
<th>Study</th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. Events</td>
<td>No. Patients</td>
<td>No. Events</td>
</tr>
<tr>
<td>Creytens 1984</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>Milo 1984</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Francois 1984</td>
<td>29</td>
<td>18</td>
</tr>
<tr>
<td>Deruttere 1987</td>
<td>42</td>
<td>31</td>
</tr>
<tr>
<td>Hannon 1987</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>Roesch 1987</td>
<td>44</td>
<td>17</td>
</tr>
<tr>
<td>De Nutte 1989</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>Hausken 1992</td>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td>Chung 1993</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Van Outryve 1993</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>Al-Quorain 1995</td>
<td>38</td>
<td>12</td>
</tr>
<tr>
<td>Kellow 1995</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>Yeoh 1997</td>
<td>21</td>
<td>19</td>
</tr>
</tbody>
</table>

Similar to the Lamotrigine add-on dataset, events in the Cisapride studies are also beneficial. The study sizes (the number of patients in the studies) are relatively small with higher observed event rates.

1.2.3 **Endoscopic Hemostasis**

The last two examples have larger sample sizes. The Endoscopic Hemostasis dataset contains a collection of 41 randomized controlled trials conducted between 1980 and 1989. These trials compare a new surgical therapy to an old one for bleeding peptic ulcers (Sacks et al., 1990; Efron, 1996; Sidik and Jonkman, 2008). Events in this dataset refer to a harmful event: recurrent bleeding.
This dataset is chosen because it was an example presented in Sidik and Jonkman (2008). We compare Sidik’s approach to a new method in Chapter 3. The complete dataset is in Appendix A.

1.2.4 Aprotinin

The last example is the largest of the four. It contains a collection of 64 randomized controlled trials on the efficacy of Aprotinin, an inhibitor used to reduce perioperative bleeding and the need for blood transfusion during cardiac surgeries (Fergusson et al., 2005). The event (perioperative transfusion) in this example is also harmful, as in the last example. Substantial heterogeneity is present in this dataset. The range of the observed event rates in both groups is from 0 to 1. The complete dataset is provided in Appendix B.

1.3 Problems to be Addressed

In medical research, many outcomes are binary indicators of a beneficial (e.g. recovery from a disease) or harmful (e.g. adverse events) event. We restrict our attention to the meta-analysis of studies with binary events in this dissertation.

Using the Lamotrigine add-on dataset as an example, we illustrate a few questions raised when synthesizing a collection of studies with binary outcomes. The estimated odds ratio and a 95% confidence interval for the odds ratio in each study are presented in Figure 1.1. Of the 11 studies, Douchowny 1999 (OR=3.82, CI=[1.96,7.45]) and Naritoku 2007 (OR=2.18, CI=[1.26,3.79]) show a significant effect of the treatment; the Jawad 1989 study (OR=11, CI=[1.06,114.09]) suggests a positive effect, but is more on the borderline with a much wider confidence interval. All other studies show
Figure 1.1: Odds ratios and 95% confidence intervals for the Lamotrigine add-on studies

statistically insignificant results at the 95% confidence level. Such a situation raises several questions:

- Is the odds ratio a good summary? Would the directions or significance of the results be different if using another measure of effect?

- How should one combine the different findings for an overall estimate of the treatment effect?

- What is the meaning of the overall treatment effect?

- What caused the differences between the studies: fundamental differences in the studies or simply chance variation?
What can we learn from these studies about a future study on the efficacy of Lamotrigine add-on treatment?

In this dissertation, we develop a flexible meta-analytical framework to better address some of these questions, focusing on appropriate models, the choice of an appropriate measure, the estimation of a meaningful overall effect (on a study population), and the prediction of the treatment effect of a future study or for a potential population.

In the first part of the work, we investigate a conditional approach to combining studies that eliminates some nuisance parameters. We show that the approach is convenient, but it neither agrees with other conditioning procedures nor captures the structure of the data as a full joint model with meaningful parameters. In the second part, we extend the full joint model with a more flexible modeling structure. Instead of forcing an \textit{a priori} decision among the common summary measures or a pre-fixed model, the flexible structure allows the data to decide which measure or model is the “best” for combining the studies. We aim to look for the best way of describing the probability distribution for studies across a two-dimensional space (probability of an event in the control group and probability of an event in the treatment group).

1.4 Organization of This Dissertation

The rest of this dissertation is organized as follows. In Chapter 2, we begin with an introduction to the problems associated with meta-analysis and a literature review of the current methods. Chapter 3 compares two different conditioning procedures and a full joint model without conditioning. A flexible modeling framework is proposed
in Chapter 4 and is extended to a Bayesian hierarchical model in Chapter 5. Chapter 6 summarizes the research and outlines the direction of future research.
Chapter 2: Literature Review

In this chapter, we review the problems associated with a meta-analysis combining findings across studies with binary events and their current solutions. The main difficulty in integrating studies is that the studies, although addressing a common question, are often different in numerous ways: the designs, the populations addressed, the treatments investigated, the outcomes recorded and so forth. The variation in the underlying effects caused by such diversity in the studies is referred to as heterogeneity. Identifying and quantifying the heterogeneity is important as it determines what assumptions should be made for combining the studies (random-effect or fixed-effect) and what measure of effect should be chosen. Different methods of combining the studies are also reviewed, including two-stage methods, likelihood-based methods, and Bayesian methods.

2.1 Heterogeneity

Investigating the causes of the heterogeneity and correctly accounting for it in the statistical analysis are critical in meta-analysis (Higgins et al., 2002). Heterogeneity can be evaluated graphically and quantitatively. Graphical tools include forest plots (e.g., Figure 1.1), Baujat plots (Baujat et al., 2002), and L’Abbé plots (Song, 1999). Quantifying the extent of heterogeneity is a more interesting statistical issue. An
obvious measure is the estimate of the between-study variance of the effects, which is part of a random-effect model estimation. However, the between-study variance estimates depend on the types of outcome (e.g., binary or continuous). The interpretation of the between-study variance varies with the measure of effect (e.g., odds ratio or risk ratio). Furthermore, direct comparison of the between-study variances of multiple meta-analyses is not possible.

Cochran’s $\chi^2$ test, a large-sample test based on the Q-statistic (defined in equation (2.7)), is a common way of testing heterogeneity of the treatment effect across studies (Whitehead and Whitehead, 1991). This test has poor power when the number of studies is small, which is a common situation in meta-analysis (Hardy and Thompson, 1998), and it doesn’t allow between-meta-analysis comparisons.

To enable the comparison of heterogeneity between meta-analyses or to quantify the impact of heterogeneity on the conclusions of a meta-analysis, Higgins and Thompson (2002) proposed some standardized statistics. The H-statistic

$$H^2 = \frac{Q}{k - 1}$$

scales the Q-statistic with the number of studies ($k$) in a meta-analysis; and

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2},$$

measures the proportion of the variation in estimates that is explained by between-study heterogeneity. $\hat{\tau}^2$ is the estimated between-study variance, and $\hat{\sigma}^2$ accounts for the within-study variance.

Both $\tau^2$ and $\sigma^2$ need to be estimated from the data. We review the estimation of $\tau^2$ in section 2.3.2. The overall within-study variance $\sigma^2$ can be estimated in different ways. For example, Takkouche et al. (1999) proposed the use of the reciprocal of
the mean weights. The weight for study $i$ is defined as $w_i = 1/\hat{\sigma}_i^2$, where $\hat{\sigma}_i^2$ is the standard error of the study-specific treatment effect. Higgins and Thompson (2002) proposed using

$$\hat{\sigma}^2 = \frac{\sum w_i (k - 1)}{(\sum w_i)^2 - \sum w_i^2},$$  \hspace{1cm} (2.1)

which leads to a relationship between $I^2$ and $H^2$

$$I^2 = \frac{H^2 - 1}{H^2}.$$ 

$I^2$ is easy to interpret and it does not depend on the type of outcomes or the number of studies as the between-study variance and the Q-statistic do.

In the presence of unexplained heterogeneity, a random-effect model, which assumes the true effect of each study is different and follows some distribution, is usually specified in a meta-analysis. There has long been debate over the fixed-effect (assumes that all studies have the same true treatment effect) and random-effect assumptions. The logic underlying fixed-effect models is that inferences are not about any general population, but instead the particular studies that were observed. Researchers with the fixed-effect point of view are generally against the idea of generalizing the inferences to a universe of studies that have not been observed and are hard to be clearly specified. This argument can be traced back to the more general argument over conditional (Fisher’s exact test) and unconditional (Pearson’s Chi-squared test) tests of $2 \times 2$ tables between Fisher and Yates (Camilli, 1990). Some very careful thought has been put into this topic (Barnard, 1947; Yates, 1984). Although this debate may never be resolved definitively, a general agreement is that it is almost always reasonable to believe that there are some between-study variations (Normand, 1999), and that the assumption of homogeneity made in fixed-effect models can rarely be satisfied.
in practice, especially for studies in the medical and social sciences. Unless there is strong evidence against it, the random-effect assumption seems tenable. Higgins et al. (2009) gave a thorough discussion on the justification and interpretation of random-effect models covering the topics of estimation, prediction, hypothesis testing, and choice of flexible distributions. They considered not only inference on the mean, but also on the whole distribution and they recommended Bayesian approaches due to their ability to capture the full uncertainty in estimating the parameters, especially for the purpose of prediction.

2.2 Choice of a Measure of the Effect

Statistical heterogeneity is defined as the variation in the true treatment effects of interest in the studies. How to parametrize the effect is usually one of the first decisions to make in many meta-analytical methods. For studies with binary outcomes (yes/no), odds ratio, risk ratio, and risk difference are commonly used. The first two measures are on a relative scale, and are typically analyzed on the logarithm scale (log odds ratio and log risk ratio). The risk difference, in contrast, is on an absolute scale.

The choice between these common measures is not trivial and has been under debate in the literature. The three major arguments are on the grounds of mathematical properties, ease of interpretation, and consistency (Deeks, 2002). Odds ratio is preferred by many from the mathematical and modeling perspective that it can take any non-negative value and on the logarithmic scale it is unbounded in both directions. Both risk ratio and risk difference have constraints: risk ratio is bounded
above depending on the control group risk, and risk difference is bounded between 
\(-1\) and 1.

Another mathematical property often considered is the symmetry of the measures. Both odds ratio and risk difference are symmetric when the coding of the event and non-event is switched, but this is not the case for risk ratio (Cox and Snell, 1989). For example, if we are interested in the mortality rate in a clinical trial, ‘death’ would be the event of interest; if the clinical trial is about a rescuing medicine to very ill patients, then ‘survival’ would be the event of interest. For risk difference, the switch of event coding will merely change the sign of the treatment effect; for the odds ratio, the new treatment effect will be the reciprocal of the original one. However, for risk ratio, the change of outcome can make a difference in the size of the treatment effect, its significance, and the heterogeneity across studies. To overcome the asymmetry problem of the risk ratio, Deeks (2002) considered using two risk ratios for each study in a meta-analysis: risk ratio of beneficial and harmful outcomes: RR(B) and RR(H). We refer to this pair of measures as risk ratio and reversed risk ratio in Chapter 4. The reversed risk ratio is the risk ratio when the original definition of event/non-event is reversed.

Despite having the best mathematical properties, the odds ratio is considered the least intuitive and the hardest to interpret. Many practitioners prefer the use of risk ratio and risk difference because they have clear clinical interpretations.

Walter (2000) argued that the choice of a measure of effect should be based on the evidence. As different measures cannot be equivalent summaries of the treatment effect (when there is an effect present), the measure that gives the best fit of the
observed data should be used. However, for most meta-analyses, there are not enough trials to compare the goodness-of-fit of different measures.

Deeks (2002) interpreted a measure fitting the data better to be “empirical evidence that one particular assumption (constant relative odds or constant relative risk or constant risk difference) is more tenable for predicting future treatment benefit of that particular intervention than the others”. Therefore, he proposed choosing a measurement that is most likely to be constant across varying baseline risk (i.e., most consistent). “Consistency” is defined as the stability of the treatment effect over different patient groups, despite variations in the baseline risk. For example, the assumption of a constant risk difference is more restrictive than the assumption of a constant relative scale measure, since the effect size varies with the baseline event rate. The same $-2\%$ risk difference may not be significant for a change from $70\%$ to $68\%$, but it may have a significant clinical meaning for a reduction from $3\%$ to $1\%$.

One way to compare which measure is more likely to be consistent on average is to test the heterogeneity of different summary statistics across many meta-analyses. Engels et al. (2000) compared the heterogeneity of odds ratio and risk ratio in 125 meta-analyses and concluded that risk difference is more likely to be heterogeneous across studies than odds ratio. Deeks (2002) carried out two complementary investigations, based on 551 meta-analyses and 114 meta-analyses respectively, on the heterogeneity of the two risk ratios $RR(B)$ and $RR(H)$. They found that on average, the risk ratio and odds ratio are equally likely to be consistent across studies, while risk difference was less consistent than both of them in all cases. In the comparison of $RR(B)$ and $RR(H)$, the latter is more consistent in most cases, especially when the intervention is to prevent adverse events.
An empirical investigation of a collection of meta-analyses provides general guidelines on the choice of measurement: choose RR(H) or odds ratio for preventive interventions and avoid risk difference for therapeutic interventions. However, it doesn’t help identify which measurement should be used for a particular meta-analysis. One approach proposed finding the least variable summary metric by selecting the one yielding the smallest heterogeneity statistic, such as the Q-statistic or $I^2$.

Such an approach may not be suitable for a few reasons. First, $I^2$ increases as the precision increases (Higgins and Thompson, 2002). The metric that gives the smallest overall standard error tends to receive the largest $I^2$, and becomes least preferred. Secondly, the studies are treated differently in the Q-statistic calculation when using different metrics (Deeks, 2002). For example, Deeks (2002) found, through examples, that using Mantel-Haenszel (MH) weights (section ), odds ratio and RR(B) give very little weight to studies with low baseline risk; and RR(H) gives almost 20 times more weight to those studies. Weights assigned by the inverse variance method are even more extreme. A metric may give a lower Q-statistic not because the summary statistics are less heterogeneous, but because it assigns lower weight to outlying trials.

We performed Cochran’s $\chi^2$ test on different measures for two datasets (Aprotinin and Lamotrigine add-on) to illustrate that for the same dataset, different measures result in different conclusions of heterogeneity. The results are presented in Appendix C.

The purpose of Deeks (2002) choosing a most consistent measure is to make predictions to a future study assuming that measure to be constant. If an inconsistent measure is chosen, the prediction would only be correct at the average baseline risk
and any extrapolation would be inappropriate. Deeks (2002) started to look in the
direction of good modeling in meta-analysis. In this paper, he deals with the issue
where if the odds ratios are normally distributed around some “true” average, then
the risk differences are not. He produced prediction patterns of the treatment event
rate given the control event rate under four different constant assumptions: constant
odds ratio, risk ratio of event, risk ratio of no event, and risk difference. The predic-
tions are all bounded in Deeks’ analysis. Using our proposed flexible measure, we are
able to make predictions without a constant measure assumption, and the predictions
from our models are not bounded (Chapter 4).

2.3 Two-Stage Methods

We have reviewed how to quantify the heterogeneity and how to choose a measure
of the effect; we now review the methods for actually combining the studies to estimate
an overall effect. The most basic and commonly used meta-analysis approach is often
described as a “two-stage” approach: the first stage is to analyse each study separately
for a point estimate (summary statistic) of a certain treatment effect of interest; and
the second stage is to pool the point estimates together, often with a weighted average
procedure. Two-stage methods can be classified into one of two categories based on
different assumptions, as we discussed in 2.1: fixed-effect and random-effect. These
two basic assumptions lead to different assignments of weights, definitions of pooled
effects, and interpretations of the results.
Table 2.1: The observed $2 \times 2$ table for study $i$

<table>
<thead>
<tr>
<th></th>
<th>events</th>
<th>no events</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>$r_{ti}$</td>
<td>$n_{ti} - r_{ti}$</td>
<td>$n_{ti}$</td>
</tr>
<tr>
<td>control</td>
<td>$r_{ci}$</td>
<td>$n_{ci} - r_{ci}$</td>
<td>$n_{ci}$</td>
</tr>
<tr>
<td>total</td>
<td>$r_{i}$</td>
<td>$n_{i} - r_{i}$</td>
<td>$n_{i}$</td>
</tr>
</tbody>
</table>

2.3.1 Fixed-Effect Methods

Fixed-effect models assume all studies included in a meta-analysis share a common true treatment effect. The observed treatment effect of study $i$ can be denoted as

$$\hat{\theta}_i = \theta + \epsilon_i,$$  \hspace{1cm} (2.2)

where $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma_i^2$.

The variance of $\epsilon_i$ ($\sigma_i^2$) is usually assumed to be known (estimated with sampling variance $\hat{\sigma}_i^2$ from the $i^{th}$ study). When $\hat{\theta}_i$ represents different measures, the sampling variances are estimated differently. For example, if $\hat{\theta}_i$ is the log odds ratio

$$\left( \frac{\log \frac{r_{ti}}{(n_{ti} - r_{ti})}}{\log \frac{r_{ci}}{(n_{ci} - r_{ci})}} \right)$$

where $r_{ti}$, $n_{ti}$, $r_{ci}$, and $n_{ci}$ are defined in Table 2.1, $\sigma_i^2$ can be estimated by

$$\hat{\sigma}_i^2 = \frac{1}{r_{ti}} + \frac{1}{(n_{ti} - r_{ti})} + \frac{1}{r_{ci}} + \frac{1}{(n_{ci} - r_{ci})}$$  \hspace{1cm} (2.3)

based on the large-sample approximation.

Under fixed-effect models, the pooled effect is an estimate of the common treatment effect $\theta$, and $\sigma_i^2$ captures the only source of uncertainty: sampling of subjects into the studies. Common fixed-effect methods include the inverse-weighted method, the Mantel-Haenszel method (Mantel and Haenszel, 1959) and Peto methods (Yusuf et al., 1985).
Inverse-Weighted Method

The weighting mechanism using study-specific $w_i = 1/\hat{\sigma}_i^2$ is called the inverse variance-weighted method (I-V). Since within-study variance is the only source of variability, it seems reasonable to give more weight to studies with higher precision (smaller variance) and less weight otherwise. The pooled effect is

$$\hat{\theta}_{I-V} = \frac{\sum_i w_i \hat{\theta}_i}{\sum_i w_i} \quad (2.4)$$

with a variance estimator

$$Var(\hat{\theta}_{I-V}) = \frac{1}{\sum_i w_i},$$

where the sums are over all studies $i = 1, 2, \ldots, k$.

Mantel-Haenszel Method Pooling Odds Ratio

The Mantel-Haenszel method was first proposed by Mantel and Haenszel to find a summary estimate of exposure effect (odds ratio) stratified by confounding factors or multiple sources in retrospective studies (Mannocci, 2009).

When applied in meta-analysis, each study is considered as a stratum. The pooled odds ratio across studies is calculated with

$$\hat{\theta}_{MH} = \frac{\sum_i r_{ti} (n_{ci} - r_{ci})/n_i}{\sum_i (n_{ti} - r_{ti}) r_{ci}/n_i}.$$ 

The MH estimator $\hat{\theta}_{MH}$ is also a weighted average of study-specific odds ratios, provided that neither of $(n_{ti} - r_{ti})$ or $r_{ci}$ is zero, with weights

$$w_{MH_i} = \frac{(n_{ti} - r_{ti})r_{ci}}{n_i}.$$ 

Under the assumption of a constant odds ratio ($\theta_i = \theta$ or $\tau^2 = 0$), the Mantel-Haenszel statistic is optimal for the test of $H_0 : \theta = 1$ (DerSimonian and Laird, 1986).
Peto Method Pooling Odds Ratio

Peto's method was first proposed by Yusuf et al. (1985) based on the MH method in their monumental paper applying meta-analysis in medical research. The pooled log odds ratio is

$$\hat{\theta}_{\text{Peto}} = \frac{\sum_i (O_i - E_i)}{\sum_i V_i},$$

(2.5)

where \(O_i = r_{ti}\) (the observed number of events in the study \(i\) treatment group), \(E_i = r_i \frac{n_{ti}}{n_i}\) (the expected number of events in the study \(i\) treatment group), and \(V_i = \frac{r_i n_{ti}(n_i - r_i) n_{ci}}{n_i^2 (n_i - 1)}\). The variance estimator is

$$\text{Var}(\hat{\theta}_{\text{Peto}}) = \frac{1}{\sum_i V_i}.$$

The Peto estimator is asymptotically unbiased under the null hypothesis of \(\theta = 1\), but it is extremely biased under unbalanced designs, or when there is a large effect under a balanced design (Greenland and Salvan, 1990).

2.3.2 Random-Effect Methods

Random-effect models allow between-study variation to capture underlying differences of the studies. Studies are considered as a random sample from a population of studies with different treatment effects, and the pooled effect size is an estimate of a hyper-parameter describing that population. The observed treatment effect of study \(i\) under a random-effect assumption can be denoted as

$$\hat{\theta}_i = \theta_i + \epsilon_i,$$

(2.6)

where \(\theta_i\) is the true treatment effect of study \(i\) with mean \(\theta\) and variance \(\tau^2\); and \(\epsilon_i\) is the sampling error of study \(i\), which is a random variable with mean 0 and
variance $\sigma_i^2$. The random-effect estimators of the overall treatment effect are also a weighted average of the individual summary statistics. The between-study variances are incorporated into the weights. As a result, the treatment effect from a large study and a small study will be assigned weights with a smaller difference, compared to the weight difference assigned by a fixed-effect model.

**DerSimonian and Laird Method**

The DerSimonian and Laird method is the simplest and most commonly used random-effect model for meta-analysis. It employs the classic form (2.6) of a random-effect model as described above. DerSimonian and Laird proposed a moment approximation procedure to estimate $\tau^2$ based on a large-sample test of heterogeneity using Cochran's $Q$-statistic

$$Q = \sum_i w_i (\hat{\theta}_i - \hat{\theta}_{I-V})^2, \quad (2.7)$$

where $\hat{\theta}_{I-V}$ is the inverse variance-weighted estimator (2.4) and $w_i$ is the reciprocal of the $i^{th}$ sampling variance. Setting $E(Q) = Q$, an estimator of $\tau^2$ is

$$\hat{\tau}^2 = \begin{cases} 
\frac{Q - (k-1)}{\sum w_i - \frac{1}{\sum w_i}} & Q > k - 1 \\
0 & Q \leq k - 1.
\end{cases} \quad (2.8)$$

$\hat{\tau}^2$ can then be used in the inverse variance weights

$$\hat{w}_{DLi} = \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2)}$$

and the DL estimator of the overall effect

$$\hat{\theta}_{DL} = \frac{\sum_{i=1}^k \hat{w}_{DLi} \hat{\theta}_i}{\sum_{i=1}^k \hat{w}_{DLi}}. \quad (2.9)$$

Both theoretical and practical examinations have shown the DL estimator to be biased. The biasedness of the DL estimator is usually considered to be a consequence
of its poor estimation of $\tau^2$, the between-study variance. Alternative estimators of $\tau^2$ have been proposed, e.g., Hunter-Schmidt, Sidik-Jonkman, Hedges, and empirical Bayes estimators (Hedges and Cooper, 1993), but none of them seem to be applied widely in practice (Sidik and Jonkman, 2008). The DL moment estimator is the default estimator in most statistical packages.

Shuster (2010) showed that random-effect meta-analysis methods using empirical weights are all biased, regardless of $\hat{\tau}^2$, unless the weights ($\hat{w}_i = (\hat{\sigma}_i^2 + \hat{\tau}^2)^{-1}$) and the individual summary statistics ($\hat{\theta}_i$) are independent. The empirical nature of the weights makes this independence hard to achieve. Our own work confirming this dependency through simulation is shown in Appendix D.

Practically, there are also good reasons to expect the presence of such correlations in medical studies. An example given by Shuster (2010) is that, in drug development, the early smaller studies of efficacy are more likely to be pure, which means they only reflect the efficacy of the drug versus the placebo. In the later, larger studies, the drug is more likely to be used in an adjuvant way (in addition to other treatments), which may cause the larger studies to show smaller effects. An opposite correlation will arise for side-effect endpoints, since the interaction between an adjuvant therapy and the drug may result in a larger effect.

### 2.3.3 Numerical Problems for Two-Stage Methods with Rare Events

Rare events (sparse data) refer to the situation when one or both arms (an arm refers to a patient group in clinical trials) have zero events. Rare events may cause numerical problems in meta-analysis using the two-stage methods.
Studies with rare events are very common in safety evaluation, a growing application of meta-analysis in medical research. Endpoints of interest in safety investigations are adverse events (AEs), the undesirable or unintended events that are associated with the use of a treatment (e.g., side effects or even deaths). Unlike more common events (usually related to efficacy or tolerability) that are well-defined during the design of a clinical trial, adverse events are usually rare and can hardly be investigated through a single study. The datasets in which AEs are not rare often come from the stage of open-label extensions of RCTs or non-randomized follow-ups after a drug is approved.

A meta-analysis of RCTs with rare events are worth special consideration, for it provides more power in detecting a relationship between an intervention and a certain type of adverse events in an environment free of confounding.

Some special challenges arise when using common two-stage methods to combine studies. First, the two-stage methods are developed by large-sample approximations and may not be suitable for small studies or studies with fewer events. Also, the relative measures (e.g., odds ratio, risk ratio) for binary data may not be defined.

One of the solutions to the problem of undefined relative measures is to add a continuity correction factor $c$ to each cell in a observed $2 \times 2$ table of a study (shown in Table 2.2). Results of a meta-analysis are usually not influenced much by the choice of $c$ if the proportion of studies with rare events is small. Otherwise, $c$ should be chosen carefully since it may bias the estimators, lower the power, and lower the coverage.

Sankey et al. (1996) compared the standard continuity correction with $c = 0.5$ to uncorrected MH and DL methods pooling odds ratios and risk differences in the
Table 2.2: An observed $2 \times 2$ table of study $i$ with continuity correction.

<table>
<thead>
<tr>
<th></th>
<th>events</th>
<th>no events</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatment</td>
<td>$r_{ti} + c$</td>
<td>$(n_{ti} - r_{ti}) + c$</td>
<td>$n_{ti} + 2c$</td>
</tr>
<tr>
<td>control</td>
<td>$r_{ci} + c$</td>
<td>$(n_{ci} - r_{ci}) + c$</td>
<td>$n_{ci} + 2c$</td>
</tr>
<tr>
<td>total</td>
<td>$r_{i} + 2c$</td>
<td>$(n_{i} - r_{i}) + 2c$</td>
<td>$n_{i} + 4c$</td>
</tr>
</tbody>
</table>

meta-analysis of sparse data, and recommended the use of a continuity correction. Adding $c = 0.5$ gives the least biased estimator of the log odds ratio for a single study (Cox, 1970), but there exist other correction methods that work better in the meta-analysis context.

Sweeting et al. (2004) explored alternative choices of $c$ for fixed-effect meta-analysis models and proposed “treatment arm correction” and “empirical correction” methods. The treatment arm correction method uses a correction factor proportional to the reciprocal of the opposite arm size. Using the study shown in Table 2.1 as an example, treatment correction method adjusts the treatment group (both events and no events) by $c/n_{ci}$ and control group by $c/n_{ti}$, where $c$ determines the magnitude of the correction. Performance of the treatment arm correction depends on the ratio of the group sizes, and it tends to pull the estimate towards ‘no effect’ (OR=1).

The empirical correction uses correction factors calculated based on the pooled effect (e.g. odds ratio) of the non-zero studies. They compared the performance of different fixed-effect pooling methods (e.g., the MH method and Peto method) and recommended using either of the new correction methods instead of $c = 0.5$. The choice between the two proposed corrections depends on the group imbalance and the pooling method used. Studies with both zero arms are excluded from the Sweeting et al.
(2004) analyses. Sweeting et al. argue that zero total event studies do not contribute to a fixed-effect meta-analysis.

The performance of the fixed-effect methods in the presence of rare events has been compared in literature. The MH estimator has better properties for sparse data (Yusuf et al., 1985), and is recommended to be used with an appropriate continuity correction when there is no heterogeneity present (Bradburn et al., 2007; Sweeting et al., 2004). In Peto’s method, studies with both zero arms do not contribute to the pooled odds ratio and its variance because the values of $O_i$, $E_i$ and $V_i$ (defined in equation (2.5)) are equal to zero. Studies with one zero arm are not affected. Bradburn et al. (2007) showed through simulation that Peto’s estimator is the least biased and most powerful with the best confidence interval coverage when the data is sparse (event rates below 1%) with little group imbalance (similar sample sizes).

Bhaumik et al. (2012) extended the work of selecting a continuity correction to random-effect meta-analysis. Similar to the fixed-effect methods, the presence of rare events biases the random-effect moment-based estimators, and the extent of the bias is proportional to the rarity. Bhaumik et al. (2012) proposed using a simple, instead of weighted, average to pool the individual effects (odds ratio) with a correction factor $c = 0.5$. They showed that such a continuity corrector results in an unbiased estimator of the overall effect up to the order of $n^{-1}$. Besides estimating the overall effect, they also proposed a new estimator for the between-study variance $\tau^2$ and a bootstrapping test for heterogeneity. Regarding the studies with no events, Bhaumik et al. (2012) concluded that including the zero-event studies with a 0.5 continuity correction helps in the estimation of the overall effect, and excluding them helps in the estimation of the heterogeneity parameter $\tau^2$. 
2.4 Likelihood-based Methods

In this section, we introduce some of the existing likelihood-based models used in meta-analysis. A focus will be put on random-effect models since we believe that it is not reasonable to assume constant parameters across studies as in the fixed-effect assumptions.

2.4.1 Approximate Models

One standard approximate likelihood-based method assumes a normal-normal (NN) structure of the observed log odds ratio $\hat{\theta}_i$

\[
\hat{\theta}_i | \theta_i \sim iid \ N(\theta_i, \sigma_i^2) \tag{2.10}
\]

\[
\theta_i \sim iid \ N(\theta, \tau^2).
\]

Under this model, $\tau^2$ can be estimated by restricted maximum likelihood (DerSimonian and Laird, 1986). $\sigma_i^2$ is assumed to be known as in the moment-based DL method. The uncertainty in estimating $\sigma_i^2$ is ignored. The estimator for the overall effect $\theta$ employs the same form as the weighted mean (2.9). Also, the NN model can be easily extended to account for covariates by assuming $\theta = X'\beta$ and to model multivariate outcomes through a multivariate normal assumption.

Despite the advantage, problems of the approximate within-study normal assumption are apparent. It shares some of the problems that the two-stage methods have. First, the effect measure should be carefully chosen to satisfy the normal assumption. Besides the logit transformations of observed rates ($\hat{\pi}_l$ and $\hat{\pi}_c$) for the odds ratio, transformation such as probit, log(-log) or arcsine could be used (Hamza et al., 2008). Second, the normal approximation is not appropriate when the sample size
is small or when the number of events is sparse. Third, the correlation between the weights \( w_i \) and \( \hat{\theta}_i \) would still cause the estimator of the overall effect to be biased. Fourth, this model still requires a continuity correction when zero arms are present.

### 2.4.2 Full Models

The within-study normal approximation can be easily avoided by directly modeling the observed counts \((n_{ti}, n_{ci}, r_{ti}, r_{ci})\) in a \(2 \times 2\) table. The four counts are the total number of patients in the treatment and control groups and the observed number of events in the treatment and control group, respectively. Such an ‘exact’ likelihood-based approach can not only avoid large-sample approximations and arbitrary continuity corrections, but more importantly, it can capture the main structure of the data with meaningful parameters.

#### Binomial-Normal Model for Odds Ratio

van Houwelingen et al. (1993) first proposed a bivariate random-effect model assuming the numbers of events in both groups of a study follow binomial distributions:

\[
\begin{align*}
  r_{ti} &\sim Bin(n_{ti}, \pi_{ti}) \\
  r_{ci} &\sim Bin(n_{ci}, \pi_{ci}).
\end{align*}
\]

Letting \( \theta_{ti} = \log[\pi_{ti}/(1-\pi_{ti})] \) (log odds in the treatment group) and \( \theta_{ci} = \log[\pi_{ci}/(1-\pi_{ci})] \) (log odds in the control group), each study produces a full likelihood function of \( \theta_i = (\theta_{ti}, \theta_{ci})' \):

\[
L(\theta_i) = \left( \begin{array}{c} n_{ti} \\ r_{ti} \end{array} \right) \left( \begin{array}{c} n_{ci} \\ r_{ci} \end{array} \right) \frac{e^{r_{ti} \theta_{ti} + r_{ci} \theta_{ci}}}{(1 + e^{\theta_{ti}})^{n_{ti}}(1 + e^{\theta_{ci}})^{n_{ci}}}
\]

The authors made a parametric random-effect assumption that \( \theta_i \) follows a bivariate normal distribution \( G = N(\mu, \Sigma) \) (binomial-normal model). The likelihood function
of $\mu$ and $\Sigma$ is written as

$$L(\mu, \Sigma) = \int L(\theta_i)Gd\theta_i.$$ 

The binomial-normal (BN) model is essentially a logistic regression model with random intercept and random slope. In extended simulation studies carried out by Hamza et al. (2008), the BN model is shown to perform consistently better than the NN model with unbiased estimates and good coverage.

The EM algorithm was used by van Houwelingen et al. to estimate the maximum likelihood estimates of $\mu$ and $\Sigma$. ‘Approximate’ estimation methods were also proposed since the ‘exact’ method requires the approximation of a two-dimensional integral which is very time consuming (van Houwelingen et al., 1993).

To construct confidence intervals for $\mu$ and $\Sigma$, Hamza et al. (2008) used the standard Wald method. Stijnen et al. (2010) used t-distributions instead of the standard normal distribution, following Proc NLMIXED in SAS, because the uncertainty in estimating the within-study variance $\sigma^2$ is not accounted for in the standard Wald intervals. van Houwelingen et al. (1993) and Hardy and Thompson (1998) constructed confidence intervals based on the likelihood ratio test and profile likelihoods.

An alternative to the binomial-normal model is a binomial-beta (BB) model (Chu et al., 2012). Instead of modeling the transformed risks (e.g. $\theta_{ti}$ and $\theta_{ci}$), the BB model assumes $\pi_i = (\pi_{ti}, \pi_{ci})'$ follows a bivariate beta distribution proposed by Sarmanov (1966).
Non-central Hypergeometric-Normal Model for Odds Ratio

The conditional non-central hypergeometric model was first proposed by van Houwelingen et al. (1993) as a likelihood extension to the Mantel-Haenszel procedure, using the conditional distribution of \( r_{ci} \) given \( r_{ti} + r_{ci} \). The likelihood function of the parameters \( w \) of each trial under a random-effect model is given by

\[
L_i(G) = \int L_i(w) dG(w),
\]

where \( w \) is a vector of \( w_i = \theta_{ci} - \theta_{ti} \), \( L_i(w) \) is a non-central hypergeometric distribution, and \( G \) is the random-effect distribution. Results from a non-parametric EM estimation of \( G \) and a parametric assumption of \( G = N(\mu, \tau^2) \) were both presented, with confidence intervals constructed from the likelihood ratio test.

This model was adopted by Sidik and Jonkman (2008) with slight modifications, mostly in the building of confidence intervals. One attractive property of the non-central hypergeometric (NH) model is that it does not have any nuisance parameters, which means the control group parameter \( \theta_{ci} \) is conditioned out.

Poisson-Gamma-Normal Model for Risk Ratio

For rare events meta-analysis, Cai et al. (2010) proposed a model based on the Poisson distribution using the risk ratio as the effect measure. \( r_{ti} \) and \( r_{ci} \) are considered independent and follow Poisson distributions with parameters \( n_{ti} \pi_{ti} \) and \( n_{ci} \pi_{ci} \) respectively. The event rate follows a log linear model

\[
\log(\pi_{ci}) = \xi_i,
\]

\[
\log(\pi_{ti}) = \xi_i + \theta_i,
\]
where $\theta_i$, the log risk ratio, measures the treatment effect and $\xi_i$ reflects the baseline risk. Instead of using a normal random-effect model for $\xi_i$, Cai et al. (2010) chose $e^{\xi_i} \sim \text{Gamma}(\alpha, \beta)$, since this gives a closed-form likelihood when $\theta_i$ is assumed to be constant. When assumed to be random, $\theta_i$ follows a normal distribution. One problem with the Gamma random-effect model is that $e^{\xi_i}$ is constrained between 0 and 1, so the values of $\alpha$ and $\beta$ should be carefully chosen, or a restricted likelihood procedure should be taken for estimating the parameters.

To avoid complicated computation, Cai et al. (2010) also proposed a conditional beta-binomial procedure with a closed-form likelihood:

\[
\begin{align*}
    r_{ti} | r_i & \sim \text{Bin}(r_i, p_i) \\
    p_i & \sim \text{Beta}(\psi \gamma, \psi W_i),
\end{align*}
\]

where $r_i = r_{ti} + r_{ci}$, $p_i = e^{\theta_i} / (W_i + e^{\theta_i})$, and $W_i = n_{ci} / n_{ti}$. $\psi$ and $\gamma$ control the center and the spread of the random-effect model, respectively. The conditional model was not preferred by the authors since the estimator was shown to be unstable with large bias in their simulation studies and it is generally harder to interpret the results from a conditional model.

**Extensions and Practice**

More flexible distributions have been explored to capture certain characteristics of the random effects. For example, a t-distribution or a skewed distribution could accommodate large estimated effects that would be considered outliers in a normal model (Lee and Thompson, 2008; Smith et al., 1995); and a mixture distribution can account for unknown groupings of the studies (Böhning, 1999).
The bivariate random-effect models can be extended to multivariate cases when there are more than two groups to be compared. An example of trivariate meta-analysis can be found in Stijnen et al. (2010). Also, the likelihood-based model can be easily applied when there are additional covariates and to analyze individual patient data (Cai et al., 2010). For example, covariates that influence the baseline risk or treatment risk can be added as a linear term; and an interaction effect could capture the influence of the covariates on the treatment effect (log odds ratio). Individual patients’ covariates, such as gender and age, can be incorporated in a similar way. A large collection of studies are required to assess influence of the covariates. Another application of the model has been demonstrated in the combination of mixed treatment comparisons, known as ‘network meta-analysis’ (Lu and Ades, 2004). Instead of comparing treatment A and B in all studies, mixed treatment comparisons includes studies comparing treatments A and B, A and C, B and C, for example.

In practice, many prefer the use of two-stage or approximation methods simply because they are the default in most meta-analysis packages. Computation involved in the likelihood-based models can be complicated, since for most of them there is no closed-form likelihood function and numerical approximations are required (e.g. Monte Carlo integration, Laplace approximation, Gaussian Quadrature and Markov Chain Monte Carlo). However, many of the likelihood methods introduced in this section can be treated as special cases of generalized linear mixed models (GLMM) and therefore estimated using standard statistical software such as NLMIXED in SAS, GLLAMM in STATA and lme4 in R. Example code for some of the likelihood-based methods is provided in van Houwelingen et al. (2002) and Stijnen et al. (2010).
2.5 Bayesian Methods

Bayesian methods are becoming more widely applied in health care research, including meta-analysis (Spiegelhalter et al., 2004). In a Bayesian analysis, both the data and the parameters are treated as random variables. Since the parameters are considered to be random quantities, prior distributions need to be specified, based on \textit{a priori} beliefs or information from evidence external to the studies in a meta-analysis. For example, a strong overall treatment effect from a meta-analysis of a collection of small trials is usually considered ‘too good to be true’. This belief can be formally incorporated into a Bayesian approach by using a prior distribution that gives less density to larger effects. The approaches above are usually referred to as fully Bayesian analyses. If the hyper-parameters in the prior distributions are estimated from the data, the approaches are termed as empirical Bayesian analyses.

Although much of the increase in applications of Bayesian methods is due to the advances in software for evaluating the Bayesian models (e.g., WinBUGS), Bayesian methods offer some potential advantages over classical approaches. They overcome many of the difficulties encountered by the traditional meta-analytical methods and easily allow for full uncertainty in all quantities, and offer natural means of interpreting the results and predicting a future study. Information from external trials can be incorporated through informative priors (Higgins and Whitehead, 1996), which is especially useful when only a small number of trials are included in a meta-analysis. A full list of the advantages of the fully Bayesian meta-analysis is presented in Chapter 8 of Spiegelhalter et al. (2004). They also described the Bayesian extension to the traditional meta-analysis models, including the two-stage models and the likelihood-based models. One thing worth noting is that many of these problems can also be
resolved using classical approaches (O’Rourke and Altman, 2005), but Bayesian approaches provide more flexibility, especially with the use of simulation-based Markov Chain Monte Carlo (MCMC) methods.

Criticisms of Bayesian meta-analyses mostly target their reliance on prior distributions, especially subjective informative ones. In response to such criticisms, many tend to choose non-informative prior distributions so that they have little or no influence on the inferences. However, the naive use of non-informative priors is usually inappropriate, especially when there is a small number of studies or when the true variance between studies is small (Smith et al., 1995; Lambert et al., 2005). Lambert et al. (2005) showed through simulation that there are usually fewer problems related to the location parameters (e.g., the overall effect) in Bayesian analyses. However, a non-informative prior is a problem for inference on the between-study variance, since the inference is potentially sensitive to the choice of priors. A sensitivity analysis of different realistic priors and a transparent report of all priors considered are always recommended. As Higgins and Spiegelhalter (2002) pointed out, ‘scepticism is a reasonable position and can be formally examined’.

Some advances in the fully Bayesian meta-analysis with informative prior distributions include:

- two-stage methods and random-effect approximate likelihood methods related to DL methodology with an emphasis on the assessment of uncertainties in meta-analytic conclusions (Carlin, 1992);

- random-effect exact likelihood methods on the log odds scale with proper prior distributions and sensitivity analysis (Smith et al., 1995);
• formulating the prior distributions for the heterogeneity parameter using data from other meta-analyses in the same therapeutic area (Higgins and Whitehead, 1996);

• extending from the log odds scale analysis to both the absolute and risk ratio scales (Warn et al., 2002);

• a re-analysis of a controversial meta-analysis containing both mega studies and small studies, demonstrating the skepticism and the choice of prior distributions (Higgins and Spiegelhalter, 2002);

• a simulation study on the choice of prior distributions for the scale parameter (Lambert et al., 2005);

• evaluating the influence of control-group risk on the heterogeneity among studies, and placing prior distributions on unknown correlation coefficients (Abrams et al., 2005);

• a reassessment of the FDA’s meta-analysis on adolescence suicidality with a hierarchical Bayesian meta-analytical approach (Kaizar et al., 2006).

2.6 Summary

There are various decisions that need to be made when carrying out a meta-analysis, for example, how should the treatment effect be parameterized, what assumptions should made regarding the studies (random effect or fixed effect), and how to actually combine the studies. These decisions influence the results and their interpretations.
For meta-analysis with binary events, the existing research is limited to choosing one of the common measures (odds ratio, risk ratio, reversed risk ratio, and relative risk) as a summary statistic for the treatment effect of a study. This choice is typically based on the ease of interpretation, the mathematical properties, and the consistency. We propose a flexible measure that allows us to estimate a measure that fits the observed data.

The moment-based estimators have been shown to be problematic in the literature. Alternatively, likelihood-based approaches, which require model specification, can be taken. We examine the current random-effect hierarchical models and propose new methods for hierarchical model selection.
The exact likelihood models we reviewed in Section 2.4.2 are built based on the observed counts in the $2 \times 2$ tables summarizing the results of randomized studies, rather than only on the relative summary measures. Such procedures are attractive as they avoid the large-sample approximation and arbitrary continuity corrections, estimate the treatment effect and the between-study variance simultaneously, and capture the main structure of the data with meaningful parameters. Despite the advantages, one difficulty that arises when modeling the table counts directly is that the distribution of each set of table counts has two parameters (typically one measures the treatment effect and the other one measures the baseline risk), rather than only one treatment effect parameter in the summary measure model. The extra baseline parameter (e.g., the log odds in control group, $\phi_c$) is often considered a nuisance parameter. Common procedures to make inferences in the presence of nuisance parameters include modeling them (e.g., van Houwelingen et al. (1993) and Smith et al. (1995)), profiling (e.g., Böhning (1999)), and conditioning on convenient statistics (e.g., Sidik and Jonkman (2008)). This last approach is the focus of this chapter.

Sidik and Jonkman (SJ) took a conditioning-based approach to meta-analysis that reduces the number of parameters to be modeled to one per study. In their
model, they condition the product binomial likelihood function on the total number of events across both groups, resulting in a non-central hypergeometric distribution. The approach captures the between-study variation through a second-level random-effect model on the treatment effect, the log odds ratio. Details of this model are introduced in Sections 3.1.2 and 3.2.2.

Intuitively, the SJ approach may not be the best one as it ignores the correlation between the treatment effect and the baseline risk. We develop a conditional approach that models the treatment effect (the log odds ratio) from a two-dimensional perspective (the full conditional model) and compare it to the SJ procedure.

The full conditional model is derived from a joint hierarchical bivariate normal model (the joint model) similar to that presented in van Houwelingen et al. (1993). We show that even though both approaches use hierarchical ideas, the SJ likelihood is not a special case of the full conditional model. As such, we prefer the joint model with naturally interpretable parameters and transparent model assumptions. Unfortunately, conditioning in the joint model does not eliminate the nuisance parameters or simplify computation, as it does for the SJ model. Therefore, we recommend the use of the joint model without conditioning, as implemented in many statistical computing packages.

3.1 Models for an Individual Study

Before considering how to combine multiple studies in a meta-analysis, we start with models for an individual randomized controlled trial with binary outcomes. For binary endpoints, each patient’s outcome is recorded as ‘yes’ if the event of interest occurred, and ‘no’ otherwise. Therefore, each study can be summarized by a $2 \times 2$
contingency table (as shown in Table 2.1). In this section, the subscript $i$ is dropped since we only consider one study.

3.1.1 A Joint Model

Patients within each study are considered independent with constant arm-specific probabilities of events ($\pi_t$ for the treatment arm and $\pi_c$ for the control arm). Then, the number of events in each group ($r_t$ and $r_c$) follows a binomial distribution

$$r_t|n_t, \pi_t \sim \text{Bin}(n_t, \pi_t)$$

$$r_c|n_c, \pi_c \sim \text{Bin}(n_c, \pi_c).$$

Further assuming that the numbers of events in the two arms are conditionally independent, their joint distribution is

$$P(r_t, r_c|n_t, n_c, \pi_t, \pi_c) = \binom{n_t}{r_t} \binom{n_c}{r_c} \pi_t^{r_t} (1 - \pi_t)^{n_t - r_t} \pi_c^{r_c} (1 - \pi_c)^{n_c - r_c}.$$ (3.2)

Following the SJ model, we intend to measure the treatment effect with the log odds ratio. We thus reparameterize the product binomial distribution as in usual logistic regression:

$$\text{logit}(\pi_j) = \phi_c + \theta X_j,$$ (3.3)

where $X_j$ is an indicator variable such that $X_j = 1$ for the treatment group ($j = 1$) and $X_j = 0$ for the control group ($j = 2$). The parameter $\phi_c$ is the log odds in the control group, or the baseline risk, and $\theta$ is the log odds ratio, the treatment effect that we are interested in estimating. The joint distribution (3.2) is thus equivalent to the likelihood function of $\phi_c$ and $\theta$:

$$L(\theta, \phi_c|n_t, n_c, r_t, r_c) = \binom{n_t}{r_t} \binom{n_c}{r_c} e^{r_t \theta} (1 + e^{\phi_c+\theta})^{-n_t} e^{(r_t+r_c)\phi_c} (1 + e^{\phi_c})^{-n_c}.$$ (3.4)
3.1.2 A Conditional Model

Because our interest lies in estimating the treatment effect of intervention $\theta$, one might turn to a conditional distribution for the events that does not depend on the underlying risk $\phi_c$. With some loss of information, $\phi_c$ can be eliminated from the likelihood by conditioning on $r = r_t + r_c$, the total number of events observed in a study. The conditional distribution of $r_t$ given the total observed number of events is non-central hypergeometric:

$$P(r_t | r., n_t, n_c, \theta) = \frac{\binom{n_t}{r_t} \binom{n_c}{r_r \cdot r_t} \exp(\theta r_t)}{\sum_{r_t \in S} \binom{n_t}{r_t} \binom{n_c}{r_r \cdot r_t} \exp(\theta r_t)}$$

(3.5)

where $S = \{ r_t : \max\{0, r. - n_c\} \leq r_t \leq \min\{n_t, r.r\} \}$. Note that for $r_t = r_c = 0$, $S = 0$, and thus $P(r_t | r., n_t, n_c, \theta)$ does not depend on $\theta$. This means that studies with two zero arms do not contribute to the estimate of the overall treatment effect under this model.

3.2 Models for Multiple Studies

In a meta-analysis, we observe multiple ($k$) independent studies and are interested in learning about the “overall” treatment effect, so assumptions on the relationships among the individual baseline risk ($\phi_{ci}$) and treatment effect ($\theta_i$) are necessary. In this section, we consider models for combining $k$ individual $2 \times 2$ tables.

3.2.1 Joint Random-Effect Models

Studies in a meta-analysis are different in many ways, such as the study design and the patient population. Due to the varying nature of the studies, we consider a
random-effect joint model to capture the potential heterogeneity:

\[ r_{ti}|n_{ti}, \pi_{ti} \sim Bin(n_{ti}, \pi_{ti}) \]

\[ r_{ci}|n_{ci}, \pi_{ci} \sim Bin(n_{ci}, \pi_{ci}) \]

\[
\logit \pi_{ji} = \phi_{ci} + \theta_i X_{ji},
\]

(3.6)

\[
\begin{pmatrix} \phi_{ci} \\ \theta_i \end{pmatrix} \overset{\text{indep}}{\sim} N \left( \begin{pmatrix} \phi_c \\ \theta \end{pmatrix}, \Sigma \right),
\]

where

\[
\Sigma = \begin{pmatrix} \omega^2 & \rho \omega \tau \\ \rho \omega \tau & \tau^2 \end{pmatrix}.
\]

In addition to the between-study heterogeneity, this model also allows correlation (\(\rho\)) between the random control group log odds (\(\phi_{ci}\)) and log odds ratio (\(\theta_i\)), which can reflect the potentially important relationship between the underlying risk and the treatment effect (Thompson et al., 1997).

As the variance components in this joint random-effect model take special values (0 or \(\infty\)), the model reduces to some of those typically considered for meta-analysis.

- \(\tau^2 = 0\) and \(\omega^2 = 0\)

\[
\logit \pi_{ij} = \phi_c + \theta X_{ij},
\]

(3.7)

This model allows no variation across studies in either of the parameters, which is identical to the case of collapsing the separate 2 \(\times\) 2 tables into one and may suffer from Simpson's Paradox (Simpson, 1951), or the less extreme amalgamation paradox (Good and Mittal, 1987).

- \(\tau^2 = 0\) and \(\omega^2 \neq 0\)

\[
\logit \pi_{ij} = \phi_{ci} + \theta X_{ij},
\]

(3.8)
This model assumes heterogeneity for the baseline risk and homogeneity for the treatment effect. $\omega^2$ captures the variation in the control group that is not explained by the sampling error. This implies that any extra observed variation in the measured outcomes is caused by the varying underlying baseline risks. The resulting analysis is the same as a fixed-effect meta-analysis model. If $\omega^2$ approaches $\infty$, we have an alternative fixed-effect model, which allows for different but fixed $\phi_{ci}$. Unfortunately, as the number of studies grows, the number of parameters grows accordingly, which may cause the estimator of the treatment effect parameter to be inconsistent (van Houwelingen and Senn, 1999).

- $\tau^2 \neq 0$ and $\omega^2 = 0$

$$\logit \pi_{ij} = \phi_c + \theta_i X_{ij}, \quad (3.9)$$

$$\theta_i \sim N(\theta, \tau^2)$$

Mirroring the idea of the model in (3.8), this model assumes the underlying risk to be fixed, while the treatment effect is heterogeneous across studies. This model corresponds to the situation when there is no significant difference among patients recruited for different studies, but the treatment effect varies. It is hard to believe this assumption would hold in a typical real application.

Re-Parametrizations of the Joint Random-Effect Model

It might be tempting to assume a priori that $\phi_{ci}$ and $\theta_i$ are independent, but this has implications on other pairs of parameters (e.g., the log odds in control group $\phi_{ci}$
and the log odds in treatment group $\phi_{ti}$ being dependent. Since $\theta_i = \phi_{ti} - \phi_{ci}$,

$$
\begin{pmatrix}
\phi_{ci} \\
\phi_{ti}
\end{pmatrix} =
\begin{pmatrix}
1 & 0 \\
1 & 1
\end{pmatrix}
\begin{pmatrix}
\phi_{ci} \\
\theta_i
\end{pmatrix},
$$

and

$$
\Sigma_{\phi_{ci}, \phi_{ti}} = 
\begin{pmatrix}
1 & 0 \\
1 & 1
\end{pmatrix}
\Sigma
\begin{pmatrix}
1 & 1 \\
0 & 1
\end{pmatrix}
= 
\begin{pmatrix}
\omega^2 & \rho\omega + \omega^2 \\
\rho\omega + \omega^2 & \omega^2 + 2\rho\omega + \tau^2
\end{pmatrix}.
$$

The correlation of $\phi_{ci}$ and $\phi_{ti}$ is therefore

$$
\rho_{\phi_{ci}, \phi_{ti}} = \frac{\rho\tau + \omega}{\sqrt{\omega^2 + 2\rho\omega + \tau^2}}.
$$

We see from equation (3.10) that if the treatment group rate ($\pi_{ti}$) is independent of the baseline rate ($\pi_{ci}$), which is equivalent to independence of the log odds in the treatment group $\phi_{ti}$ and the log odds in the control group $\phi_{ci}$, we have $\rho_{\phi_{ci}, \phi_{ti}} = 0$. This implies the correlation between the baseline log odds $\phi_{ci}$ and the treatment effect $\theta_i$ (the log odds ratio) is

$$
\rho = -\frac{\omega}{\tau};
$$

if the treatment effect is independent of the baseline log odds ($\rho = 0$), we calculate

$$
\rho_{\phi_{ci}, \phi_{ti}} = \frac{\omega}{\sqrt{\omega^2 + \tau^2}}.
$$

Another pair of parameters that we may be interested in is the log odds ratio $\theta_i$ and the average log odds $\frac{\phi_{ti} + \phi_{ci}}{2}$. The variance-covariance matrix for these parameters is

$$
\Sigma_{\theta_i, \frac{\phi_{ti} + \phi_{ci}}{2}} = 
\begin{pmatrix}
\tau^2 & \frac{\tau^2}{2} + \rho\omega \\
\frac{\tau^2}{2} + \rho\omega & \frac{\tau^2}{4} + \rho\omega + \omega^2
\end{pmatrix}.
$$
\( \rho_{\theta_i, \phi_{ti} + \phi_{ci}} = 0 \) implies
\[
\rho = -\frac{\tau}{2\omega};
\] (3.12)
and \( \rho = 0 \) implies
\[
\rho_{\theta_i, (\phi_{ti} + \phi_{ci})/2} = \frac{\tau}{\tau^2 + 4\omega^2}.
\]

### 3.2.2 The Non-central Hypergeometric Conditional Model

To combine the studies based on the individual-study conditional distributions (3.5), Sidik and Jonkman (2008) propose allowing study-to-study heterogeneity in the treatment effect \( \theta \) via the conditional model in Section 3.1.2:

\[
r_{ti}|r_{-i}, n_{ti}, n_{ci}, \theta_i \text{ indep} \sim \text{Non-central hypergeometric}(n_{ti}, n_{ci}, r_{-i}, \theta_i) \quad (3.13)
\]
\[
\theta_i \text{ indep} \sim N(\theta, \tau^2).
\]

The likelihood function of \( \theta \) and \( \tau^2 \) for \( k \) independent studies under this model is
\[
L^{NCH}(\theta, \tau^2) = \prod_{i=1}^{k} L_i^{NCH}(\theta, \tau^2) = \prod_{i=1}^{k} \int L_i^{NCH}(\theta_i) \frac{1}{\tau} \varphi \left( \frac{\theta_i - \theta}{\tau} \right) d\theta_i,
\] (3.14)
where \( L_i^{NCH}(\theta_i) = P(r_{ti}|n_{ti}, r_{ci}, r_{-i}, \theta_i) \) (3.5), and \( \varphi(\cdot) \) is the standard normal density function.

### 3.2.3 A Full Conditional Model

The method proposed by Sidik and Jonkman (2008) has a hierarchical feel to it, in which the hierarchy is constructed on the conditional distribution (3.5) for each study. This is not the only way, however, to build a conditional hierarchical model for this problem. In this section, we consider another approach to conditioning by first constructing a joint distribution (3.6) for all the study table counts, and then
finding the conditional distribution by conditioning on the vector of event totals for all of the studies, the same statistics conditioned on in the SJ model.

The likelihood function for this procedure can be constructed as follows:

\textbf{Step 1.} For study $i$, one can find $P(r_{ti}, r_{ci}|n_{ti}, n_{ci}, \theta, \phi_{ci})$, derived from $P(r_{ti}, r_{ci}|n_{ti}, n_{ci}, \theta, \phi_{ci})$, the product of two binomial probability mass functions. Based on the full model (3.6), we can find

$$P(r_{ti}, r_{ci}|n_{ti}, n_{ci}, \theta, \phi_{ci}, \theta, \phi_{ci}, \tau^2, \omega^2, \rho)$$

where $\pi(\theta, \phi_{ci}|\theta, \phi_{ci}, \tau^2, \omega^2, \rho)$ is the density of the bivariate normal distribution.

\textbf{Step 2.} Conditioning on $r_{ci}$, we have

$$P(r_{ti}|r_{ci}, n_{ti}, n_{ci}, \theta, \phi_{ci}, \tau^2, \omega^2, \rho)$$

where $S_i = \{r_{ti} : max\{0, r_{.} - n_2\} \leq r_{ti} \leq min\{n_1, r_{.}\}\}$, as in the non-central hypergeometric likelihood function (3.14).

\textbf{Step 3.} The likelihood function of $\theta = (\theta, \phi_{ci}, \tau^2, \omega^2, \rho)$ for the collection of studies is

$$L_{FC}(\theta) = \prod_i L_{FC}^i(\theta)$$

(3.16)
We call this conditioning approach the full conditional model, since it is derived from a full likelihood-based model.

The point of using a conditional model (the NCH hierarchical model) was to eliminate nuisance parameters, however, when we construct a conditional model from a full joint model, it does not appear that the nuisance parameter $\phi_c$ can be eliminated. The likelihood function (3.16) is not equivalent to the non-central hypergeometric likelihood function (3.14) (except in trivial special cases). However, we are interested in learning whether the two conditional probability distributions result in the same estimators for the overall treatment effect ($\theta$) and between-study variance $\tau^2$. Because neither of the two models gives a closed-form likelihood function, using numerical approximation methods is necessary.

3.3 Approximating the Likelihood Functions and Estimating the Parameters

To approximate the non-central hypergeometric model likelihood function $L_i^{NCH}(\theta, \tau^2)$ as in (3.14), we applied a grid-based integration method. A grid of $\theta_i$ spanning from $-15$ to 15 with a grid width 0.05 was used. Denoting the $g^{th}$ element in the grid as $\theta_{i,g}$, we calculated a grid of values for $L_i^{NCH}(\theta, \tau^2)$. For any fixed pair ($\theta, \tau^2$), we calculated the normal density $w_{i,g} = \frac{1}{\tau} \varphi\left(\frac{\theta_{i,g} - \theta}{\tau}\right)$, and approximated the integral.

$$L_i^{NCH}(\theta, \tau^2) = \int L_i^{NCH}(\theta_i) \frac{1}{\tau} \varphi\left(\frac{\theta_i - \theta}{\tau}\right) d\theta_i \approx \frac{\sum_{g} L_i^{NCH}(\theta_{i,g}) w_{i,g}}{\sum_{g} w_{i,g}}.$$ (3.17)

Although there exist more advanced numerical methods, the grid method described above is shown empirically to be computationally efficient and adequate for this approximation. To decide the grid width, we plotted the evaluation results (3.17) and the corresponding computation time for different grid widths in Figure 3.1. We see
that the numerical result stabilizes when the grid width is smaller than 0.5, and the computation time starts to increase substantially around 0.1. Considering both accuracy and efficiency, we picked a grid width of 0.05. A similar investigation for Monte Carlo simulation (results not shown) demonstrated that basic Monte Carlo methods are much more time consuming to achieve an equivalent level of accuracy.

The approximation of the full conditional model likelihood function (3.15) is more complicated since it involves a double integral in the bivariate normal random-effect model. In this case, we chose the Monte Carlo integration method instead of grid evaluation as above because the two-dimensional grid method is not as efficient. A sample of \((\theta_i, \phi_{ci})\) of size \(n_{MC} = 50000\) is simulated from the bivariate normal distribution \(\pi(\theta_i, \phi_{ci}|\theta, \phi_c, \tau^2, \omega^2, \rho)\) given a known combination of \((\theta, \phi_c, \tau^2, \omega^2, \rho)\). Denoting the \(m^{th}\) pair as \((\theta_{i,m}, \phi_{ci,m})\), we have the double integral in \(L_{FC}^i\) (3.15) approximated as

\[
\int \int P(r_{ti}, r_{ci}|n_{ti}, n_{ci}, \theta_i, \phi_{ci})\pi(\theta_i, \phi_{ci}|\theta, \phi_c, \tau^2, \omega^2, \rho)d\theta_i d\phi_{ci} \\
\approx \frac{1}{n_{MC}} \sum_m P(r_{ti}, r_{ci}|n_{ti}, n_{ci}, \theta_{i,m}, \phi_{ci,m}).
\]

We estimate the parameters using the maximum likelihood method. One difficulty using Monte Carlo integration in this case is that the numerically approximated likelihood function is not smooth for finding a maximum. We thus fit a smoothing spline ANOVA model (Gu, 2013) to the approximated \(L_{FC}^i(\theta)\) to find the maximum likelihood estimates of the parameters.

Likelihood-based confidence intervals are constructed based on the profile log likelihood functions of \(\theta\) and \(\tau^2\). Hence the 95% marginal confidence intervals for the two parameters are given by all values of \(\theta\) and \(\tau^2\) that satisfy

\[
l_\theta^*(\theta) > l^*(\hat{\theta}) - 3.84/2
\]
Figure 3.1: Grid evaluation of $L_{i}^{NCH}(\theta, \tau^2)$ and the corresponding time used at different grid widths.
and
\[ l^*_r(\tau^2) > l^*(\hat{\tau}^2) - 3.84/2, \]
where 3.84 is the 95% critical value of the \( \chi^2_1 \) distribution.

### 3.4 Illustration and Comparisons

In this section, we apply the two conditional procedures to two datasets to illustrate the differences in the likelihood functions and in the estimates of the log odds ratio \( \theta \) and between-study variance \( \tau^2 \). As discussed above, the numerical approximations of the full conditional model (3.15) can be computationally intensive, especially when the dataset is large, and when the vector of the total number of events in the studies is not sparse. With this in mind, we first considered a simple artificial dataset containing only one study. In Section 3.4.2, we present results from a real dataset with 39 studies.

#### 3.4.1 A Numerical Example

The purpose of this example is simply to show that the likelihood functions of the two procedures, when evaluated at the same values of the parameters, give different results. In this example, we let the numbers of patients for the treatment group and the control group both be 50, and the number of events in both groups be one (i.e., \( n_t = n_c = 50, r_t = r_c = 1 \) and \( r. = 2 \)). The Monte Carlo integration method (sample size 50,000) was used to approximate the likelihood functions at \( \theta = (-3, -1, 0, 1, 3) \), \( \omega^2 = (0.02, 1, 2, 3) \) and \( \rho = (-0.9, -0.5, 0, 0.5, 0.9) \). \( \phi_c \) was fixed to be 0 and \( \tau^2 \) to be 1. The results of the numerical example are summarized in the plots in Figure 3.2.
Figure 3.2: Comparison of the non-central hypergeometric likelihood and the full conditional model likelihood functions at given parameter values.

From Figure 3.2a, we see that the hypergeometric likelihood function does not change with $\omega^2$ and $\rho$, which is not surprising since the nuisance parameters are conditioned out in this model. We also see that the likelihood function is symmetric about $\theta = 0$ at the chosen values of $\theta$ ($-3$, $-1$, $1$, and $3$). On the other hand, the full conditional model likelihood function (3.16) shown in Figure 3.2b is unimodal with respect to $\rho$ where the location of the mode increases with $\theta$. The pattern in the likelihood function with respect to $\omega^2$ is less straightforward to see from Figure 3.2b, but it is very similar to that of $\rho$. These results confirm that the two likelihood functions are not the same, and that the nuisance parameters can not be conditioned out by conditioning on the number of events in the full conditional model.
3.4.2 Endoscopic Hemostasis Data

We illustrate the difference in the two approaches using the endoscopic hemostasis dataset described in Section 1.2.3 as in Sidik and Jonkman (2008). Studies 40 and 41 are excluded from the analysis for the reasons that study 40 is considered to be an “overly influential outlier” and that study 41 does not contribute to the overall likelihood (Efron, 1996).

![Graph of profile log likelihoods of θ and τ² for the endoscopic hemostasis dataset.](image)

Figure 3.3: The profile log likelihood functions of θ and τ² for the endoscopic hemostasis dataset. The horizontal dashed lines are the maximum log likelihood values minus 1.92, indicating the confidence intervals. The vertical dotted lines show the MLEs. The non-central hypergeometric method is denoted by NHG and the full conditional model by FCOND.

The profile log likelihood functions of θ and τ² of both models are shown in Figure 3.3. The horizontal dashed lines are the maximum log likelihood values minus 1.92,
Table 3.1: The maximum likelihood estimates of $\theta$ and $\tau^2$ and the 95% confidence intervals using the non-central hypergeometric (NHG) and full conditional (FCOND) models.

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimates of $\theta$</th>
<th>Estimates of $\tau^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\theta}$</td>
<td>CI</td>
</tr>
<tr>
<td>NHG</td>
<td>-1.20</td>
<td>(-1.64, -0.76)</td>
</tr>
<tr>
<td>FCOND</td>
<td>-1.41</td>
<td>(-1.70, -1.14)</td>
</tr>
</tbody>
</table>

indicating the corresponding confidence intervals. The vertical dotted lines show the MLEs. The estimates and the confidence intervals are shown in Table 3.1.

Figure 3.3 confirms that the two conditioning approaches are different and that they result in different estimates of the parameters. Figure 3.3a shows that the profile log likelihood function of $\theta$ is flatter under the non-central hypergeometric model than under the full conditional model, which results in a wider confidence interval for $\hat{\theta}_{NHG}$. Both models estimate the overall log odds ratio ($\theta$) to be significantly smaller than 0 (no effect), which means the new surgical therapy effectively reduces the occurrence of recurrent bleeding. The full conditional model estimates a larger negative effect ($-1.41$, with a confidence interval $(-1.70, -1.14)$) than the NHG model does ($-1.20$, with a confidence interval $(-1.64, -0.76)$). The profile log likelihood functions of $\tau^2$ for the two models (Figure 3.3b) are also different. $\hat{\tau}^2_{NHG} = 1.25$ (0.53, 2.77) is much smaller than $\hat{\tau}^2_{FCOND} = 2.16$ (1.21, 3.52). The widths of the two confidence intervals are not much different.
3.5 Conclusions

Sidik and Jonkman (2008) proposed a likelihood-based conditioning procedure for combining independent studies with binary events. They used the non-central hypergeometric distribution resulting from conditioning on both margins in an individual $2 \times 2$ table to estimate the overall log odds ratio and the study-to-study heterogeneity. This conditioning procedure eliminates the nuisance parameters related to the control group information, and accommodates zero studies. In Appendix E, we showed in a simulation study that the SJ model's mean effect estimator is only unbiased under certain independence assumptions, but is often less biased than the DL estimator. We also showed that the NHG between-study variance estimator is less biased compared to the DL variance estimator.

In this chapter, we considered a different conditioning procedure based on a full joint binomial-normal hierarchical model. Instead of first finding a conditional distribution for each study and then assuming a hierarchical distribution for the log odds ratios, we first find the joint hierarchical distribution for all studies and then condition on the vector of total numbers of events. Looking at the likelihood functions, and comparing results from an artificial example and a real dataset, we recognize that the two conditional approaches are different and could yield different maximum likelihood estimates. Although the non-central hypergeometric model provides a convenient way of dealing with the nuisance parameter, it is not expected to agree with the full conditional model and therefore cannot be used to estimate the same parameters. We recommend that the use of the conditional non-central hypergeometric
model as a parameter-reducing approach should be justified by more than mathematical convenience, and that a model-based procedure capturing the main structure of a problem with meaningful parameters should be the approach taken.

In summary, the conditional model should not be preferred over a full joint model for a few reasons. First of all, the conditioning does not actually remove the nuisance parameters, which was considered the primary advantage of a conditional procedure. Secondly, only studies with at least one event contribute to the conditional inference. The unconditional random-effect analysis allows all studies to contribute, and might therefore be more powerful. Thirdly, evaluating the full conditional model is computationally intensive and time consuming. It requires numerical approximation to the likelihood function containing a double integral. Finally, a full joint model can be easily extended to a fully Bayesian analysis with meaningful parameters.
Chapter 4: A Flexible Measure and Two-Dimensional Distribution for Combining Studies in Meta-Analysis of Binary Events

The existing meta-analysis procedures pre-specify a “consistent” effect measure (e.g., odds ratio, risk ratio, and risk difference). These analyses are conditional on the assumptions that the pre-specified measures and models are good choices. In this section, we develop a flexible measure of effect that is determined by the data. Using our measure, we describe flexible two-dimensional distributions for combining studies in a meta-analysis of binary events.

4.1 A Flexible Measure of Effect

Common measures of effect size in meta-analysis include the odds ratio, risk ratio and risk difference. Reversed risk ratio (the risk ratio when the original coding of an event and absence of an event are switched) is also considered by some researchers. We summarize the current research on measure selection in Section 2.2.

All common measures of effect size can be viewed as a transformation from the original parameters \( \pi_c \) and \( \pi_t \), the control and the treatment group risks (Rücker et al., 2009). Denote a transformation as \( g(\cdot) \), the measures can be written in a
general form

\[ g(\pi_t) - g(\pi_c). \]

Table 4.1 summarizes the measures, the transformations and their estimates.

Noticing the relationship between the three relative scale measures (log odds ratio \( \theta \), log risk ratio \( \eta \) and log reverse risk ratio \( \eta_R \))

\[ \theta = \eta - \eta_R, \]

we propose a new measure: a linear combination of the latter two measures weighted by an unknown parameter \( \alpha \)

\[ \theta(\alpha) = \alpha \eta - (1 - \alpha) \eta_R, \quad \alpha \in [0, 1]. \]

We consider \( \alpha \) as a transit parameter, since when it takes different values, \( \theta(\alpha) \) becomes different measures. It preserves the direction of the relationship such that increasing values of \( \theta(\alpha) \) imply increasing rates of events in the treatment group relative to the control group. When there is no treatment effect, \( \theta(\alpha) \) equals zero at any value of \( \alpha \). The three original relative measures are (scaled) special cases of \( \theta(\alpha) \) when \( \alpha \) takes the values of 0, 0.5 and 1 (shown in Table 4.2).

For an observed study, the newly proposed measure is estimated by a summary statistic

\[ \hat{\theta}(\alpha) = \alpha \hat{\eta} - (1 - \alpha) \hat{\eta}_R \]

\[ = \alpha \log \frac{r_t}{n_t} - \frac{r_c}{n_c} - (1 - \alpha) \log \frac{1 - r_t/n_t}{1 - r_c/n_c} \]

for a given value of \( \alpha \).
Table 4.1: Common measures, transformations, and estimates for a meta-analysis of binary events.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Parameter</th>
<th>Transformation</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>log odds ratio</td>
<td>$\log \frac{\pi_t}{1-\pi_t} - \log \frac{\pi_c}{1-\pi_c}$</td>
<td>logit</td>
<td>$\log \frac{r_t}{n_t-r_t} - \log \frac{r_c}{n_c-r_c}$</td>
</tr>
<tr>
<td>log risk ratio</td>
<td>$\log \pi_t - \log \pi_c$</td>
<td>log</td>
<td>$\log \frac{\pi_t}{\pi_c} - \log \frac{r_t}{r_c}$</td>
</tr>
<tr>
<td>log reversed risk ratio</td>
<td>$\log (1 - \pi_t) - \log (1 - \pi_c)$</td>
<td>-</td>
<td>$\log \frac{\pi_t}{n_t-r_t} - \log \frac{n_c-r_c}{n_c}$</td>
</tr>
<tr>
<td>risk difference</td>
<td>$\pi_t - \pi_c$</td>
<td>identity</td>
<td>$\frac{n_t-r_t}{n_c-r_c}$</td>
</tr>
</tbody>
</table>
Table 4.2: Special cases of the new summary metric $\theta(\alpha)$.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\theta(\alpha)$</th>
<th>Notation</th>
<th>Corresponding common measure</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$-\log\frac{1-\pi_t}{1-\pi_c}$</td>
<td>$-\eta_R$</td>
<td>log reversed risk ratio</td>
<td>negative</td>
</tr>
<tr>
<td>0.5</td>
<td>$\frac{1}{2}\log\frac{\pi_t/(1-\pi_t)}{\pi_c/(1-\pi_c)}$</td>
<td>$\theta/2$</td>
<td>log odds ratio</td>
<td>half</td>
</tr>
<tr>
<td>1</td>
<td>$\log\frac{\pi_t}{\pi_c}$</td>
<td>$\eta$</td>
<td>log risk ratio</td>
<td>identical</td>
</tr>
</tbody>
</table>

4.1.1 Approximate Distribution

Assuming that the number of events in each group of a study follows a binomial distribution as in (3.1), we find that $\theta(\alpha)$ follows an approximately normal distribution based on the normal approximation to the binomial distribution and the delta method:

$$\hat{\theta}(\alpha) = \alpha\hat{\eta} - (1 - \alpha)\hat{\eta}_R \approx_{\text{approx}} \text{Normal}(\mu_\alpha, \sigma^2_\alpha),$$

(4.2)

where

$$\mu_\alpha = \alpha \log\frac{\pi_t}{\pi_c} - (1 - \alpha) \log\frac{1 - \pi_t}{1 - \pi_c};$$

$$\sigma^2_\alpha = \alpha^2\left(\frac{1 - \pi_t}{n_t\pi_t} + \frac{1 - \pi_c}{n_c\pi_c}\right) + (1 - \alpha)^2\left(\frac{\pi_t}{n_t(1 - \pi_t)} + \frac{\pi_c}{n_c(1 - \pi_c)}\right) + 2\alpha(1 - \alpha)\left(\frac{1}{n_t} + \frac{1}{n_c}\right).$$

The derivation is presented in Appendix F.

The normal approximation gives an estimator of the within-study variance ($\hat{\sigma}^2$) for a given $\alpha$

$$\hat{\sigma}^2 = \alpha^2\left(\frac{1}{r_t} - \frac{1}{n_t} + \frac{1}{n_c} - \frac{1}{r_c}\right) + (1 - \alpha)^2\left(\frac{1}{n_t - r_t} - \frac{1}{n_t} + \frac{1}{n_c - r_c} - \frac{1}{n_c}\right) + 2\alpha(1 - \alpha)\left(\frac{1}{n_t} + \frac{1}{n_c}\right),$$

(4.3)
and the approximate variance estimator can be used to compute a 95% confidence interval for $\theta_i(\alpha)$:

$$\hat{\theta}_i(\alpha) \pm 1.96\hat{\sigma}_{ai}.$$  \hspace{1cm} (4.4)

Using the approximate normal distribution of $\theta_i(\alpha)$, we can write a Normal-Normal (NN) random-effect hierarchical model for summary statistics $\hat{\theta}_i(\alpha)$

$$\hat{\theta}_i(\alpha) \sim N(\theta_{ai}, \sigma_{ai}^2) \hspace{1cm} (4.5)$$

$$\theta_{ai} \sim N(\theta_\alpha, \tau_\alpha^2),$$

where $\sigma_{ai}^2$ is the variance of the observed measure $\hat{\theta}_i(\alpha)$, usually assumed to be known.

To fit this model, we can find the maximum likelihood function (ML) or restricted maximum likelihood function (REML) estimates of parameters $\theta_\alpha$ and $\tau_\alpha^2$ for a fixed value of $\alpha$. Fitting the model for a grid of $\alpha$ values between 0 and 1 gives the profile likelihood of $\alpha$ and maximizing the profile likelihood gives the maximum likelihood estimate. For each value of $\alpha$, the model can be implemented using standard statistical packages for linear hierarchical models, such as the lmer() function in R package lme4.

In the NN hierarchical model (4.5), a normality assumption is made for the study-level summary statistic $\hat{\theta}_i(\alpha)$. This model may not be appropriate when we have a small number of studies, when the events in the studies are rare, and also when $\alpha$ takes extreme values close to 0 or 1. And similar to odds ratio, $\theta(\alpha)$ is not an intuitive measure to interpret.

### 4.1.2 Illustration

The Lamotrigine add-on (Section 1.2.1) and Aprotinin (Section 1.2.4) datasets are used to show how the estimates and the confidence intervals for $\theta_i(\alpha)$ change as the value of $\alpha$ changes from 0 to 1.
To see the distributions of the studies in the two datasets, we visualized them using the L’Abbé plot, which is a graphical tool in meta-analysis (L’Abbé et al., 1987). These plots show the observed treatment group event rates against the observed control group event rates. L’Abbé plots of the two datasets are shown in Figure 4.1. The sizes of the circles are proportional to the corresponding studies’ precision (reciprocal of the estimated variance of the risk differences).

![L'Abbé plots](image)

(a) 11 trials of Lamotrigine add-on  
(b) 64 trials of Aprotinin

Figure 4.1: L’Abbé plots on the risk difference scale. Each study is represented by a circle, the size (area in 1/16 inches) of which is proportional to its precision (inverse variance of the risk difference). The diagonal line is an indication of no effect ($\pi_t = \pi_c$).

We see in Figure 4.1a that the Lamotrigine add-on studies are scattered in the bottom left corner with small observed event rates in both groups. The endpoint in the Lamotrigine add-on studies is beneficial, so the studies strictly above the diagonal line of $\pi_t = \pi_c$ show a significant positive effect of the treatment. The observed treatment
event rates seem to increase as the observed control event rates increase. The studies in the Aprotinin dataset as plotted in Figure 4.1b show a great deal of heterogeneity. The observed event rates in both groups have a range from 0 to 1, and they seem to be positively related as well. As the endpoint in this collection of studies is a harmful event (bleeding during cardiac surgeries), observations below the diagonal line are preferable.

Using the normal approximation, confidence interval (95%) plots of the two datasets on the $\theta(\alpha)$ scale for $\alpha = 0, 0.5, \text{ and } 1$ are presented in Figures 4.2 and 4.3. The studies are plotted in increasing order of observed control group event rates (i.e., study 1 has the smallest $\hat{\pi}_c$). For the Aprotinin dataset, a continuity correction of 0.5 is used when necessary in the calculation of the standard errors of $\hat{\theta}_i(\alpha)$ using formula (4.3). The confidence intervals of the Lamotrigine add-on studies are all quite narrow when $\alpha$ is 0, and they continue to grow wider as $\alpha$ increases to 1. The estimated individual mean effect $\hat{\theta}_i(\alpha)$ increases as $\alpha$ increases as well (since for all studies in this dataset, $0.5 > \hat{\pi}_{ti} > \hat{\pi}_{ci}$).

As seen in the L’Abbé plots, the range of the observed control event rates in the Aprotinin dataset is much wider compared to the studies in the Lamotrigine add-on dataset. When $\alpha$ is 0, studies with lower control risks give narrower confidence intervals than those with larger control risks, and the confidence interval plots show a “funnel” shape. When $\alpha$ is 0.5, the widths of the intervals become less diverse and as $\alpha$ approaches 1, the direction of the funnel shape is reversed, where larger control risk studies give narrower confidence intervals. The changes in $\hat{\theta}_i(\alpha)$ for this dataset is not as consistent as those in the Lamotrigine add-on dataset. For the studies with a lower observed treatment event rate than the observed control rate, which is the
case for most of the studies in the Aprotinin dataset since the outcome is harmful, \( \hat{\theta}_i(\alpha) \) decreases as \( \alpha \) increase when the observed event rates are smaller than 0.5, and vice versa when the observed event rates are greater than 0.5. Overall, we see from the two datasets that \( \alpha \) affects the studies with extreme control risks (i.e., closer to 0 or 1) more than those with moderate rates.

### 4.1.3 Quantifying the Heterogeneity

In this section, we seek to evaluate the different patterns by quantifying the heterogeneity of \( \hat{\theta}_i(\alpha) \) and finding the \( \alpha \) that minimizes the study-to-study variability on the scale of \( \theta(\alpha) \). \( \theta(\hat{\alpha}) \) can thus be considered as the most "consistent" measure across the observed studies. Common ways to measure the extent of heterogeneity between the studies are discussed in Section 2.1. We extend the Q-statistic and \( I^2 \) to the \( \theta(\alpha) \) scale and propose a new statistic to better quantify the study-to-study variation.

Following the typical treatment of summary measures, we denote an estimate of parameter \( \theta_i(\alpha) \) from study \( i \) (\( i = 1, 2, \ldots, k \)) by \( \hat{\theta}_i(\alpha) \), and its precision by \( w_{\alpha i} = \hat{\sigma}_{\alpha i}^{-2} \). \( \hat{\sigma}_{\alpha i}^2 \) is the within-study variation defined in (4.3), and is treated as known. The inverse variance estimator of an overall \( \theta(\alpha) \) is calculated as

\[
\mu_{F\alpha} = \frac{\sum w_{\alpha i} \hat{\theta}_i(\alpha)}{\sum w_{\alpha i}}, \tag{4.6}
\]

and the Q-statistic for the heterogeneity test for \( \hat{\theta}_i(\alpha) \) as

\[
Q = \sum w_{\alpha i} (\hat{\theta}_i(\alpha) - \mu_{F\alpha})^2.
\]
Figure 4.2: Confidence interval plots of the 11 Lamotrigine add-on trials on the $\theta(\alpha)$ scale for $\alpha = 0, 0.5$ and 1. The studies are ordered by increasing observed control group event rates.
Figure 4.3: Confidence interval plots of the 64 Aprotinin trials on the $\theta(\alpha)$ scale for $\alpha = 0, 0.5$ and 1. The studies are ordered by increasing observed control group event rates. A continuity correction of 0.5 is used when necessary in the standard error calculation.
Using the moment-based DL estimator (2.8), the between-study variance of $\hat{\theta}_i(\alpha)$ can be estimated by

$$\hat{\tau}_2^\alpha = \begin{cases} 
\frac{Q-(k-1)}{\sum w_{ai}-\sum w_{ai}^2} & Q > k - 1 \\
0 & Q \leq k - 1 
\end{cases};$$

and $I^2$, the proportion of the variation in estimates that is explained by the between-study heterogeneity, is therefore

$$I^2 = \frac{\hat{\tau}_2^2}{\hat{\tau}_2^2 + \hat{\sigma}_2^2}. \tag{4.7}$$

$\hat{\sigma}_2^2$ describes the “average” within-study variance of the studies, which can be calculated using Equation (2.1), as proposed by Higgins and Thompson.

An $\alpha$ that gives the smallest Q-statistic or $I^2$ can be chosen and the corresponding $\theta(\alpha)$ may be considered the most consistent measure and the “best” summary of the observed data. This procedure of choosing the $\alpha$ that yields the smallest heterogeneity statistics carries over the problems discussed in Section 2.2 that the studies are treated differently under different values of $\alpha$ and that $I^2$ increases as the precision of the measurements decreases.

Essentially, in the traditional $I^2$ calculation (4.7), for a fixed $\alpha$, all studies are considered to have the same within-study variance $\hat{\sigma}_2^2$. This means that $I^2$ actually measures the proportion of the variation explained by the between-study variance for a collection of ‘typical’ studies. It may be more attractive to measure an average variability rather than variability of a typical study, so we propose to use an average individual $I^2$ defined as

$$I^2_{AI} = \frac{1}{k} \sum I^2_i = \frac{1}{k} \sum \frac{\hat{\tau}_2^2}{\hat{\tau}_2^2 + \hat{\sigma}_2^2}. \tag{4.8}$$
Instead of treating all studies as a ‘typical’ study, the average individual $I^2$ allows different precisions for different studies. It is still a number between 0 and 1, and measures the average contribution of the between-study variation to the total variation of each study.

We calculate $I^2$ and $I_{AI}^2$ at different values of $\alpha$ for the Lamotrigine add-on (1.2.1) and the Aprotinin (1.2.4) dataset and present the results in Figure 4.4. Results from

![Graphs showing $I^2$ and $I_{AI}^2$ for Lamotrigine add-on and Aprotinin datasets.](image)

Figure 4.4: $I^2$ and $I_{AI}^2$ calculated from the Lamotrigine add-on and the Aprotinin dataset. The vertical lines show the $\alpha$ values that result in the smallest heterogeneity measured by $I^2$ and $I_{AI}^2$.

the Lamotrigine add-on dataset is less interesting, as the estimate of $\tau^2$ reduces to zero at $\alpha = 0.1$. From the Aprotinin dataset, one can see that $I_{AI}^2$ is constantly smaller than $I^2$. The pattern of $I_{AI}^2$ is flatter compared to that of $I^2$, which means
that $\tilde{I}_{\tilde{A}}^2$ is affected less by the values of $\alpha$. The value of $\alpha$ that gives the smallest $I^2$ is 0.58, and the smallest $\tilde{I}_{\tilde{A}}^2$ is given by $\alpha = 0.59$.

### 4.2 Flexible Two-Dimensional Inference

In the last section, we proposed a new effect measure and discussed how to make inference based only on the summary statistics. The methods in this section are developed to seek the “best” transformation to adequately describe the probability distribution of the observed data across the two-dimensional space of $\pi_t$ and $\pi_c$ rather than focus on a one-dimensional summary, and to make inference in a meta-analysis.

#### 4.2.1 Random-effect Hierarchical Model

We have discussed in Chapters 2 and 3 the advantages of the random-effect joint models (e.g., model (3.6)) in which we model the counts of a $2 \times 2$ table instead of a summary statistic (e.g., model (4.5)). In this section, we extend the joint model approach to a flexible modeling framework based on the new measure $\theta(\alpha)$ to describe the joint distribution of the ‘true’ risks in the treatment ($\pi_t$) and the control ($\pi_c$) groups in a meta-analysis of binary events.

By rewriting the new measure $\theta(\alpha)$ as

$$
\theta(\alpha) = \alpha \log \frac{\pi_t}{\pi_c} - (1 - \alpha) \log \frac{1 - \pi_t}{1 - \pi_c}
$$

$$
= \log \frac{\pi_t^\alpha}{(1 - \pi_t)^{1-\alpha}} - \log \frac{\pi_c^\alpha}{(1 - \pi_c)^{1-\alpha}},
$$

we recognize a transformation

$$
g_\alpha(\mu) = \log \frac{\mu^\alpha}{(1 - \mu)^{1-\alpha}}
$$

so that $\theta(\alpha) = g_\alpha(\pi_t) - g_\alpha(\pi_c)$. Using this transformation as a link function, we can incorporate the new measure $\theta(\alpha)$ in a generalized linear mixed-effect model (GLMM)
A framework. Such a link function is different from the traditional ones in the sense that it contains an unknown parameter $\alpha$. For a given value of $\alpha$, $g_{\alpha}(\cdot)$ is a one-to-one function, which is appropriate as a link function. We write the GLMM with $\alpha$ as:

- the data model

$$r_{ti}|n_{ti}, \pi_{ti} \sim Bin(n_{ti}, \pi_{ti}) \quad (4.11)$$

$$r_{ci}|n_{ci}, \pi_{ci} \sim Bin(n_{ci}, \pi_{ci})$$

- the link function

$$g_{\alpha}(\mu) = \log \frac{\mu^\alpha}{(1 - \mu)^{1-\alpha}}.$$  

- the linear predictors and the random effects

$$g_{\alpha}(\pi_{ti}) = \mu_i(\alpha) + \theta_i(\alpha) \quad (4.12)$$

$$g_{\alpha}(\pi_{ci}) = \mu_i(\alpha)$$

$$\begin{pmatrix} \mu_i(\alpha) \\ \theta_i(\alpha) \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_\alpha \\ \theta_\alpha \end{pmatrix}, \begin{pmatrix} \omega^2_\alpha & \rho_\alpha \omega_\alpha \tau_\alpha \\ \rho_\alpha \omega_\alpha \tau_\alpha & \tau^2_\alpha \end{pmatrix} \right) . \quad (4.13)$$

We refer to model (4.11-4.13) as the “flexible GLMM” or the “GLMM with $\alpha$” in the following discussion.

The traditional random-effect logistic regression model (equivalent to $\alpha = 0.5$ in equation (4.12)) assumes bivariate normality on the log odds ratio and log odds in the control group scale. A contour plot of such an assumption is illustrated in Figure 4.5. However, if the normality assumption is true on a different scale (e.g., a different value of $\alpha$), the traditional logistic regression model becomes inappropriate. We illustrate a possible situation in Figure 4.6.

Figure 4.6a presents a contour plot describing the distribution of $\mu(0.7)$ and $\theta(0.7)$, which is assumed to be normal. The contour plot is produced using two-dimensional
Figure 4.5: A contour plot of log odds ratio vs. log odds in the control group under the bivariate normal assumption. The true log odds in the control group is $-2.21$, and the true log odds ratio is $0.23$. Values of the other parameters are $\pi_t = 0.3$, $\pi_c = 0.1$, $\tau^2 = 1$, $\omega^2 = 0.5$, and $\rho = 0.9$. 
kernel density estimation from random draws $\mu_i(0.7)$ and $\theta_i(0.7)$ ($n = 10,000$), simulated from the bivariate normal distribution in (4.13). True parameters used for the simulation are $\alpha = 0.7$, $\pi_t = 0.3$, $\pi_c = 0.1$, $\tau^2_\alpha = 1$, $\omega^2_\alpha = 0.5$, and $\rho_\alpha = 0.9$. If the above assumption is true, modeling $\mu(0.5)$ and $\theta(0.5)$ with a bivariate normal distribution is inappropriate. To show this, we transform the simulated study-specific probabilities to the scale of $\mu_i(0.5)$ and $\theta_i(0.5)$ and produce the contour plot in Figure 4.6b. One can see that the normality assumption on the scale of $\mu(0.7)$ and $\theta(0.7)$ shown in Figure 4.6a broke on the scale of $\mu(0.5)$ and $\theta(0.5)$ shown in Figure 4.6b. The flexible GLMM we proposed allows us to capture such a situation and to use the data to estimate the “best” $\alpha$ (i.e., the “best” measure) to describe the probability distribution for studies across a two-dimensional space (probabilities in the treatment group and probabilities in the control group).

4.2.2 Implementation

We find the maximum likelihood estimate (MLE) of $\alpha$ by maximizing its profile log likelihood function

$$l^*(\alpha) = l(\hat{\Theta}_{[-\alpha]} | \alpha),$$  \hspace{1cm} (4.14)

where $\hat{\Theta}_{[-\alpha]}$ is the MLE for all other parameters at an arbitrary but fixed value of $\alpha$, i.e.,

$$l(\hat{\Theta}_{[-\alpha]} | \alpha) \geq l(\Theta_{[-\alpha]} | \alpha)$$

for all $\Theta_{[-\alpha]}$. A 95% large sample likelihood-based confidence interval for $\alpha$ is given by all values that satisfy

$$l^*(\alpha) > l^*(\hat{\alpha}) - 3.84/2,$$  \hspace{1cm} (4.15)
where 3.84 is the 95% critical value of the $\chi^2_1$ distribution. This is a practical way to build a confidence interval for $\alpha$, but we do not have a solid theoretical basis for this approach. Future work could study the coverage probability of this interval by simulation. We estimate the parameters $\Theta_{[-\alpha]}$ using their MLEs at $\hat{\alpha}_{MLE}$.

The profile likelihood function of $\alpha$ and the MLEs of $\Theta_{[-\alpha]}$ can be found by fitting the flexible GLMM at different values of $\alpha$. To fit the GLMMs, we used the function `glmer()` in R package lme4 (R Core Team, 2013; Bates et al., 2013). This function relates the linear predictor to the conditional mean of the response through the inverse link function. The integrals are evaluated using the default Laplace approximation, and the optimizer is Nelder-Mead.

Figure 4.6: Contour plots of $\theta(\alpha)$ vs. $\mu(\alpha)$. The normality assumption is made for $\alpha = 0.7$. The true parameters are $\pi_t = 0.3$, $\pi_c = 0.1$, $\tau_\alpha^2 = 1$, $\omega_\alpha^2 = 0.5$, and $\rho_\alpha = 0.9$. The plots are produced through simulation and kernel density estimation.
The link function (4.10) needs to be specified for function `glmer()`. Notice that neither a closed-form inverse function for \( g_\alpha(\cdot) \), nor its first derivative is available for an arbitrary \( \alpha \), but they are required by the iteratively reweighted least squares (IWLS) algorithm and need to be numerically evaluated. For a given \( \alpha \), we calculate \( \eta_{\alpha,p} = g(\alpha, p) \) for a fine grid (from \( 10^{-5} \) to \( 1 - 10^5 \) with a grid width \( 10^5 \)) of probability \( p \). Given a value of \( \eta \), the corresponding value of \( p \) can be located. Interpolation is used if the given \( \eta \) is between of two values in the grid \( \eta_{\alpha,p} \). Such a fine grid is chosen so that the IWLS algorithm converges. The first derivative of \( g_\alpha(\cdot) \) is approximated by \( \frac{\Delta \eta_{\alpha,p}}{\Delta p} \), where \( \Delta \) represents “change in”. The R code for the user specified link function and an example for fitting model (4.11-4.13) is attached in Appendix G.

Choosing a good starting point is also important for the IWLS algorithm to converge. Plus, it helps to prevent the variance components estimates from being trapped in an area on the edge of the parameter space where they are close to zero. We choose the starting point in a sequential way by first fitting the model at \( \alpha = 0.5 \) and using its estimates as the starting point for the model with the next value on the grid of \( \alpha \) (0.49 and 0.51 for example).

The MLEs of the original parameters are hard to interpret. When \( \alpha \) takes different values, the corresponding \( \theta(\alpha) \) have different meanings. Therefore, the interpretation of the parameters \( \hat{\Theta}_{|\alpha|} \) are particular to a given value of \( \alpha \). However, \( \theta(\alpha) \) is a transformations from the original scale: the event rates in the treatment and the control group. We can thus convert \( \theta(\alpha) \) and \( \mu(\alpha) \) to \( \pi_t \) and \( \pi_c \) for interpretation, where \( \pi_t \) and \( \pi_c \) are the medians of the distributions of \( \pi_{ti} \) and \( \pi_{ci} \).
4.2.3 Simulation Study

We conduct a simulation study to compare the performance of the flexible GLMM and the random-effect logistic regression model (3.6). In each simulation replication, \( k = 50 \) studies were simulated from model (4.11-4.13). The number of patients in the two arms of each study is sampled by rounding a random draw from a normal distribution with mean 100 and standard deviation 10. These numbers are chosen based on simulation studies in literature and the dataset Aprotinin.

\( \theta_i(\alpha) \) and \( \mu_i(\alpha) \) are generated from a bivariate normal distribution for given values of \( \alpha, \theta_\alpha = g(\alpha, \pi_t) - g(\alpha, \pi_c), \mu_\alpha = g(\alpha, \pi_c), \tau^2, \omega^2 \) and \( \rho \). 1000 replicate analyses were performed for each of two sets of parameter values. The true parameter values and the estimation results for the simulation study are shown in Table 4.3. At \( \alpha = 0.3, \tau^2 = 1, \omega^2 = 0.5 \) and \( \rho = -0.5 \), we have \( \pi_c > \pi_t > 0.5 \) in the first simulation and \( \pi_c < \pi_t < 0.5 \) in the second. \( \alpha \) is 0.7 in simulation 3 and 4, and the other parameters are the same as in simulation 1 and 2, respectively. In all four simulation studies, the GLMM retrieves the original parameters used to simulate the data.

\( \pi_t \) and \( \pi_c \) are estimated by the medians of the distributions of \( \pi_{ti} \) and \( \pi_{ci} \). They are comparable between the logistic model and the GLMM model since they are on the original scale without transformation. The logistic model underestimates the event rates \( \pi_t \) and \( \pi_c \) when \( \alpha = 0.3 \) and overestimates them when \( \alpha = 0.7 \).

The estimates of the variance components \( \tau^2 \) and \( \omega^2 \) are not comparable between the two models, because their interpretations are specific to the particular scales. However, the differences in the magnitude of the estimates of them (especially in the second and the third simulation) seem larger than would be explained by the difference in interpretation. It also underestimates \( \rho \) in the first and the fourth case and...
## Table 4.3: Estimation results for the simulation studies.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Model</th>
<th>Parameter Estimation (Standard Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Ture value</td>
<td>Logistic</td>
<td>0.3 1 0.5</td>
</tr>
<tr>
<td>1</td>
<td>GLMM</td>
<td>0.33(0.11) 0.93(0.26) 0.46(0.15)</td>
</tr>
<tr>
<td>2</td>
<td>Logistic</td>
<td>0.3 1 0.5</td>
</tr>
<tr>
<td>GLMM</td>
<td>0.31(0.08) 1.01(0.32) 0.52(0.19)</td>
<td>0.20(0.06) 0.10(0.03) -0.50(0.12)</td>
</tr>
<tr>
<td>3</td>
<td>Logistic</td>
<td>0.7 1 0.5</td>
</tr>
<tr>
<td>GLMM</td>
<td>0.70(0.07) 1.00(0.30) 0.51(0.17)</td>
<td>0.80(0.06) 0.90(0.03) -0.48(0.13)</td>
</tr>
<tr>
<td>4</td>
<td>Logistic</td>
<td>0.7 1 0.5</td>
</tr>
<tr>
<td>GLMM</td>
<td>0.67(0.11) 0.94(0.27) 0.46(0.15)</td>
<td>0.21(0.04) 0.10(0.01) -0.48(0.13)</td>
</tr>
</tbody>
</table>

Logistic denotes the random-effect logistic regression model and GLMM denotes the random-effect GLMM on the scale of $\theta(\alpha)$. $\pi_t$, $\pi_c$ (estimated by the medians of the distributions of $\pi_{ti}$ and $\pi_{ci}$) are the comparable parameters of the two models.
overestimates $\rho$ in the second and third case. Overall, the flexible GLMM functions well.

4.3 Prediction

One of the most important strengths of a meta-analysis is its ability to formally assess the generalizability of the findings from one collection of studies to a future study, especially with the random-effect assumptions. Predictive distributions are considered ‘potentially the most relevant and complete statistical inferences to be drawn from random-effect meta-analysis’ (Higgins et al., 2009). Methods in this section predict parameters or variables of interest in a future study using the flexible GLMM (4.11-4.13). Model (4.11-4.13) determines the probability distribution that adequately fits the data across the space of $\pi_{ti}$ and $\pi_{ci}$. Such a distribution can then be applied to predict the results of a future study.

Traditionally, a meta-analysis is used for the prediction of $\theta_{new}$, the effect in a new study (e.g., log odds ratio in the random-effect logistic regression model), to determine whether there is a significant treatment effect. Under the flexible GLMM, a similar inference is to be made for $\theta(\hat{\alpha})_{new}$, where $\hat{\alpha}$ is estimated from the data so that $\theta(\hat{\alpha})$ is the ‘best’ measure.

It is straightforward to make predictions on the original scale $\pi_t$ given the control group information $\pi_c^*$. Researchers carrying out studies usually have a rough idea what control group event rate to expect for a certain group of patients. When the patients information is not pre-specified, the predictive distribution of $\pi_t|\pi_c$ can be used to obtain a prediction band for $\pi_t$ at any value of $\pi_c \in [0, 1]$. Such a prediction
band can be used to decide for which group of patients the treatment is going to be effective.

The number of patients in each group of a new study is fixed in the study design. It is also interesting to predict the possible number of events to expect in the treatment group given \( \pi^*_c \) and \( n^*_i \), since \( \pi_i \) is a latent parameter in the binomial data structure.

Common practice in developing the predictive distribution is to treat the variance components \( \tau^2_\alpha \) and \( \omega^2_\alpha \) and the correlation \( \rho_\alpha \) as known. They are estimated in model (4.11-4.13) and the fitted model for the random effects is

\[
\begin{pmatrix}
\mu_i(\hat{\alpha}) \\
\theta_i(\hat{\alpha})
\end{pmatrix}
\sim
N
\left( \begin{pmatrix}
\hat{\mu}_\alpha \\
\hat{\theta}_\alpha
\end{pmatrix},
\begin{pmatrix}
\hat{\omega}_\alpha^2 & \hat{\rho}_\alpha \hat{\omega}_\alpha \hat{\tau}_\alpha \\
\hat{\rho}_\alpha \hat{\omega}_\alpha \hat{\tau}_\alpha & \hat{\tau}_\alpha^2
\end{pmatrix}\right).
\tag{4.16}
\]

Let \( \pi^*_c \) be the control group event rate that we want to make predictions on for a future study. \( \hat{\alpha} \) is estimated from the current meta-analysis, and

\[
\mu^*(\hat{\alpha}) = \log \frac{\pi^*_c \hat{\alpha}}{(1 - \pi^*_c)^{1-\hat{\alpha}}}
\tag{4.17}
\]

\[
\hat{\mu}_{\text{new}} = \hat{\theta}_\alpha + \frac{\tau}{\omega} \rho (\mu^*(\hat{\alpha}) - \hat{\mu}_\alpha).
\tag{4.18}
\]

The conditional distribution of \( \theta_{\text{new}}(\hat{\alpha}) \) for a new study given \( \hat{\alpha} \) and \( \pi^*_c \) is thus given by

\[
\theta(\hat{\alpha})_{\text{new}}|\mu^*(\hat{\alpha}) \sim N \left( \theta^*, (1 - \rho^2)\tau^2 \right).
\]

To incorporate the estimation uncertainty of the mean parameters, let

\[
\theta^* \sim N(\hat{\mu}_{\text{new}}, SE(\hat{\mu}_{\text{new}})^2),
\]

where

\[
SE(\hat{\mu}_{\text{new}})^2 = SE(\hat{\theta}_\alpha)^2 + \left( \frac{\tau}{\omega} \rho \right)^2 SE(\hat{\mu}_\alpha)^2 - 2 \frac{\tau}{\omega} \rho \text{Cov}(\hat{\theta}_\alpha, \hat{\mu}_\alpha). \tag{4.19}
\]
The predictive distribution for \( \theta(\hat{\alpha})_{\text{new}} = g_{\alpha}(\pi_{t,\text{new}}) = \log \frac{\pi_{t,\text{new}}^{\hat{\alpha}}}{(1 - \pi_{t,\text{new}})^{1-\hat{\alpha}}} \) is therefore

\[
\theta(\hat{\alpha})_{\text{new}} \sim N\left(\hat{\mu}_{\text{new}}, (1 - \rho^2)\tau^2 + SE(\hat{\mu}_{\text{new}})^2\right),
\]

(4.20)

and \( \pi_{t,\text{new}} \) has a probability density function (pdf)

\[
h(\pi_{t,\text{new}}) = f(\theta(\hat{\alpha})_{\text{new}}) \frac{\partial g_{\alpha}(\pi_{t,\text{new}})}{\partial \pi_{t,\text{new}}}
= f(g_{\alpha}(\pi_{t,\text{new}})) \left(\frac{\hat{\alpha}}{\pi_{t,\text{new}}} + \frac{1 - \hat{\alpha}}{1 - \pi_{t,\text{new}}}\right),
\]

(4.21)

where \( f(\cdot) \) is the pdf of the normal distribution in (4.20). Thus a 95% prediction interval for \( \theta(\alpha)_{\text{new}} \) can be obtained by

\[
\hat{\mu}_{\text{new}} \pm 1.96 \sqrt{(1 - \rho^2)\tau^2 + SE(\hat{\mu}_{\text{new}})^2},
\]

(4.22)

and that for \( \pi_{t,\text{new}} \) by a transformation

\[
g_{\alpha}^{-1}\left(\hat{\mu}_{\text{new}} \pm 1.96 \sqrt{(1 - \rho^2)\tau^2 + SE(\hat{\mu}_{\text{new}})^2}\right).
\]

(4.23)

Given a total number of patients \( (n_t^* \) for the treatment arm of a future study, the predictive distribution of \( r_{t,\text{new}} \) is given by

\[
P(r_{t,\text{new}} = y) = \int_0^1 P(r_{t,\text{new}} = y|n_t^*, \pi_{t,\text{new}}) h(\pi_{t,\text{new}}) d\pi_{t,\text{new}}
\]

(4.24)

where \( y = 0, 1, \ldots, n_t^* \) and \( P(r_{t,\text{new}} = y|n_t^*, \pi_{t,\text{new}}) \) is the probability mass function of a binomial distribution with parameters \( n_t^* \) and \( \pi_{t,\text{new}} \).

We use the Monte Carlo method to numerically evaluate the integral in (4.24). A sample of \( \theta(\hat{\alpha})_{\text{new}} \) is simulated from the normal distribution (4.20), and is transformed to a sample of \( \pi_{t,\text{new}} \) using a numerical inverse link function \( g_{\alpha}^{-1}(\theta(\hat{\alpha})_{\text{new}}) \). Denote each element in the sample as \( \pi_{t,m} \). \( P(r_{t,\text{new}} = y) \) is approximated by

\[
\frac{1}{n_{MC}} \sum_{m=1}^{n_{MC}} \binom{n_t^*}{y} \pi_{t,m}^y (1 - \pi_{t,m})^{n_t^*-y} \left(\frac{\hat{\alpha}}{\pi_{t,m}} + \frac{1 - \hat{\alpha}}{1 - \pi_{t,m}}\right),
\]

79
where $n_{MC}$ is the Monte Carlo sample size.

The construction of the predictive interval for $r_{t,new}$ is slightly more complicated since $r_{t,new}$ is a discrete random variable. We use a 95% highest probability density (HPD) interval, which is the shortest interval that covers 95% of the probability (Chen and Shao, 1999).

4.4 Examples

We apply our proposed two-dimensional methods to the four examples described in Section 1.2.

The profile log likelihood functions of $\alpha$ are shown in Figure 4.7. The dotted lines in Figures 4.7c and 4.7d indicate the 95% likelihood-based confidence interval defined in (4.15). We see that in all four datasets, the distribution of $\alpha$ is asymmetric. The likelihood functions of $\alpha$ estimated using the Lamotrigine and Cisapride datasets show little difference across the parameter space (less than 1), which means the two datasets do not provide enough information for a 95% confidence interval of $\alpha$ that is narrower than $(0, 1)$.

The estimates and their corresponding 95% confidence intervals for our proposed model can be compared with those from the random-effect logistic regression model (Table 4.4). For simplicity, we use the same notation for parameters of the two models in Table 4.4, although they have different meanings. $\theta$ and $\mu$ represent the mean parameters in the random-effect models, $\tau^2$ and $\omega^2$ are the variance components, and $\rho$ is the correlation coefficient. In the logistic model, $\theta$ is the overall log odds ratio and $\mu$ is the mean of the distribution for the log odds in the control group. In the GLMM with $\alpha$, $\theta$ and $\mu$ are the mean of the distribution of $\theta_i(\alpha)$ and $\mu_i(\alpha)$,
Figure 4.7: Profile log likelihood function of $\alpha$. The dotted line indicates the 95% likelihood-based confidence intervals defined in (4.15).
respectively. To enable comparison, we convert them back to the original scale \( \pi_t \) and \( \pi_c \), which can be interpreted as the medians of marginal distributions of \( \pi_{ti} \) and \( \pi_{ci} \).

We see in Table 4.4 that there is no difference in the estimations of \( \pi_t \) and \( \pi_c \) in the first two datasets. This is consistent with what we observed in Figure 4.7a and 4.7b. In the last two datasets, the logistic model estimates the control group rate to be higher than the GLMM does. The GLMM tends to produce shorter confidence intervals in the last two datasets as well. Another comparable parameter of the two models is \( \rho \). The extent of correlation estimated by the GLMM is consistently smaller than that estimated under the logistic model.

We notice from these four examples that for studies with harmful endpoints (Endoscopic and Aprotinin), where lower treatment event rate than the control event rate is preferred, \( \alpha \) is estimated to be larger than 0.5. For studies with beneficial endpoints (Lamotrigine add-on and Cisapride), \( \alpha \) is estimated to be smaller than 0.5. It is unclear whether there is a mechanism behind this ad hoc rationalization, but it certainly is an interesting area for further investigation.

Contour plots of the estimated distributions of \((\theta_i(\alpha), \mu_i(\alpha))\) at \( \hat{\alpha} \) (4.16) are presented in Figure 4.8c. Given information in the control group \((\pi^*_c)\) and \( \hat{\alpha} \), we calculate \( \mu^*(\hat{\alpha}) = g(\hat{\alpha}, \pi^*_c) \) (the blue vertical lines) and illustrate what the predictions for \( \theta(\hat{\alpha}) \) look like (the horizontal blue dashed lines) on the plots. For example, in Figure 4.8a, we see that the prediction interval given by the contour plot is strictly below the estimated mean for \( \theta(\hat{\alpha}) \) (the horizontal red dotted line). The values \( n^*_c \), \( \pi^*_c \), and \( n^*_t \) used for predictions and the corresponding results are presented in Table 4.5.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Parameter Estimation</th>
<th>Parameter Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lamotrigine</td>
<td>Logistic</td>
<td>$\theta$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>add-on</td>
<td></td>
<td>0.39</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>GLMM</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>Cisapride</td>
<td>Logistic</td>
<td>0.44</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>GLMM</td>
<td>0.19</td>
<td>0.42</td>
</tr>
<tr>
<td>Endoscopic</td>
<td>Logistic</td>
<td>0.76</td>
<td>-0.14</td>
</tr>
<tr>
<td>Hemostasis</td>
<td>GLMM</td>
<td>0.24</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Logistic</td>
<td>0.60</td>
<td>-0.56</td>
</tr>
<tr>
<td>Aprotinin</td>
<td>GLMM</td>
<td>0.30</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 4.4: Estimation results for the four examples. Logistic denotes the random-effect logistic regression model and GLMM denotes the random-effect GLMM on the scale of $\theta(\alpha)$.

$\hat{\alpha}, \hat{\theta}, \hat{\mu}, \hat{\omega}^2, \hat{\rho}, \hat{\pi}_c,$ and $\hat{\pi}_t$ are the comparable parameters of the two models.
Figure 4.8: Estimated predictive distributions of $(\theta_i(\alpha), \mu_i(\alpha))$ at $\hat{\alpha}$ with predictions for $\theta_{new}(\hat{\alpha})$ at given $\mu^*(\hat{\alpha})$. 
Table 4.5: Prediction results for the treatment group event rate $\pi_t$ given the control group event rate $\pi_c^*$, and the treatment group number of events $r_t$ given $n_t^*$. The known information $n_c^*$, $\pi_c^*$, and $n_t^*$ are from a chosen study in the original datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Observations</th>
<th>Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n_c^*$</td>
<td>$\pi_c^*$</td>
</tr>
<tr>
<td>Lamotrigine add-on</td>
<td>50</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cisapride</td>
<td>22</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endoscopic Hemostasis</td>
<td>74</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aprotinin</td>
<td>14</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It is more straightforward to make predictions on the original scale of the event rates $\pi_t$ and $\pi_c$. For each example, we predict $\pi_{t,new}$ for a grid of $\pi_c^*$ values from 0 to 1 with 95% prediction intervals, shown in Figures 4.9 to 4.12. The diagonal line is an indication of no effect, and the points represent the observed data. Marginal predictive distributions of $\pi_c$ and $\pi_t$ for a new study are also presented. The GLMM results are presented in red, compared to the Logistic model results in blue. The four examples show very different features.

The Lamotrigine add-on dataset (Figure 4.9) studies efficacy of the drug with a beneficial outcome, so $\pi_t > \pi_c$ is more desirable. At the given information $\pi_c^* = 0.1$ (the green vertical line), we predict $\pi_{t,new}$ to be 0.18 with a 95% prediction interval (0.13, 0.24). The prediction band falls strictly above the diagonal line of no effect, which means that we expect the Lamotrigine add-on treatment to be beneficial for all future groups of patients. The treatment effect seems very consistent across different types of patients as there is a strong positive correlation between $\pi_c$ and $\pi_t$. The
prediction band is wider for moderate values of $\pi_c$ and narrower for values close to 0 and 1. However, we see that the observed $\pi_c$ does not exceed 0.24, and any statement outside of the range of the observed data is dangerous.

The Cisapride treatment event rate does not change much for different patient groups as evidenced by the fact that the confidence band in Figure 4.10 is relatively flat. When the treatment is applied to a patient group with $\pi_c < 0.47$, we expect to see significant benefit because the prediction intervals in that range of the values are strictly above the diagonal line. For example, the predicted $\pi_{t,new}$ for $\pi_c = 0.27$ is 0.74 with a 95% prediction interval $(0.49, 0.89)$. If $\pi_c$ is between 0.47 and 0.72, we predict $\pi_t$ to be larger than $\pi_c$, but not strictly. The prediction band indicates that when the event rate in the control group is higher than 0.9, the drug is not expected to be beneficial. This is an extrapolation of the data as well and, therefore, is not a reliable prediction. For studies with beneficial events as outcomes, a control group event rate close to 1 means the placebo or the control treatment works well enough and no treatment is needed. Predictions for such a case are not practical.

The Endoscopic Hemostasis dataset prediction band shows a similar flat pattern as in the Cisapride dataset. The endpoint for this collection of studies is bleeding, a harmful event, and $\pi_c < \pi_t$ is preferable. A prediction when the control event rates are close to 0 (no bleeding occurs in this group of patients) is not helpful. We predict that the treatment works on patient groups with higher risk of bleeding (greater than 0.51).

Both the Endoscopic Hemostasis dataset and Cisapride dataset show small correlation between the treatment group event rate and the control group event rate, which may suggest a problematic data collection process in the included studies. For
example, selection criteria for the studies may be different, or the blinding may be
violated.

The Aprotinin dataset shows great heterogeneity in the observed event rates in
both groups: they both have a range from 0 to 1. The recorded endpoint is bleeding
during cardiac surgeries, a harmful outcome. Similar to the Endoscopic Hemostasis
dataset, \( \pi_t \) smaller than \( \pi_c \) is preferred. From the prediction band in Figure 4.12, we
see that a significant effect of Aprotinin is expected for patients with higher risk of
bleeding \( (\pi_c > 0.46) \). For \( \pi_c^* = 0.5 \), we predict \( \pi_{t,new} \) to be 0.27 with a 95% prediction
interval \( (0.13, 0.50) \).

Figure 4.13 shows the predictive distributions of \( r_{t,new} \) given \( n_c, \pi_c, \) and \( n_t \) cal-
culated using (4.24) and the values listed in Table 4.5. The red point in each plot
corresponds to the predicted number of events in the treatment group (highest prob-
ability). Because \( r_t \) is a discrete random variable, we are not able to find the exact
95% prediction intervals. The two intervals with confidence levels closest to 95% are
presented for each dataset. The blue and green points indicate the prediction inter-
vals with smaller and larger confidence, respectively. The exact confidence level for
each interval is noted on the plots.

4.5 Summary

In this section, we proposed a new measure \( \theta(\alpha) \), a linear combination of two
common relative measures log risk ratio and reversed log risk ratio. From a one-
dimensional perspective, we estimate a “best” summary measure fitting the observed
data by finding the \( \alpha \) that minimizes the study-to-study variation.
Figure 4.9: Lamotrigine add-on. Predictions on $\pi_t$ for different values of $\pi_c$. The points are the observed data. The black solid line represents the predicted values and the dashed red lines show the 95% prediction band. The diagonal line indicates no effect. $\pi_t > \pi_c$ is preferable.
Figure 4.10: Cisapride. Predictions on $\pi_t$ for different values of $\pi_c$. The points are the observed data. The black solid line represents the predicted values and the dashed red lines show the 95% prediction band. The diagonal line indicates no effect. $\pi_t > \pi_c$ is preferable.
Figure 4.11: Endoscopic Hemostasis. Predictions on $\pi_t$ for different values of $\pi_c$. The points are the observed data. The black solid line represents the predicted values and the dashed red lines show the 95% prediction band. The diagonal line indicates no effect. $\pi_t < \pi_c$ is preferable.
Figure 4.12: Aprotinin. Predictions on $\pi_t$ for different values of $\pi_c$. The points are the observed data. The black solid line represents the predicted values and the dashed red lines show the 95% prediction band. The diagonal line indicates no effect. $\pi_t < \pi_c$ is preferable.
Figure 4.13: Prediction for the number of events in the treatment arm \( r_t \) of a future study given the number of patients in both arms \( (n_t, n_c) \) and the control group risk \( (\pi_c) \). The blue and green points indicate two prediction intervals of \( r_t \) that are closest to the 95% confidence level.
\( \theta(\alpha) \) is also incorporated into a two-dimensional generalized linear model, with which we describe the probability distribution of the observed data across the two-dimensional space of \( \pi_t \) and \( \pi_c \) and make predictions on future studies.

The two-dimensional GLMM is a flexible extension of the random-effect likelihood-based approaches, and thus inherits the advantages of the approaches. It models the binomial structure of the data directly, captures the correlation between the control and the treatment groups, allows the investigation of heterogeneity through study-level covariates (i.e., underlying risk), and eases the prediction to a future study. An improvement of our method over the traditional random-effect models, e.g. a logistic regression or a log linear model, is that the unknown parameter \( \alpha \) allows us not to have a pre-specified model. Instead, we estimate the best model for the data. We do see from the examples that when the sample size is too small, locating the “best” \( \alpha \) could be difficult.

It has been in literature that a summary measure should be decided from the data (Walter, 2000). However, the existing methods are still limited to the choice from one of the common measures. Measure \( \theta(\alpha) \) allows an estimation the the scale that truly fit the data the best. It is not an intuitive measure, but having a measure that is straightforward to interpret is not the focus of our approach. We prefer to make statistical inference on the space across the original scale of \( \pi_t \) and \( \pi_c \). The predictions on \( \pi_t \) and \( \pi_c \) can be transformed into prediction intervals on any scale.

An important contribution of the approach we take is this chapter is that we set up a framework to think about more general distributions on \( (\pi_{ti}, \pi_{ci}) \). The proposed measure \( \theta(\alpha) \) is just one possibility. For example, instead of mixing odds ratio and risk ratio as for \( \theta(\alpha) \), we might mix odds ratio and risk difference, which will be more
appropriate for the rare events situation. Non-parametric approach can be taken as well. The approach to thinking about prediction can be readily employed for any further joint prior in the GLMM framework, and would only need small changes (if any) for even more general hierarchical models.

Furthermore, the flexible framework takes the modeling uncertainty into account. For a traditional analysis, a measure and/or a model are pre-specified. All inferences are made conditionally on the assumption that the pre-specifications are not bad and that the uncertainty in choosing a measure/model is not accounted for. The methods we build in this section acknowledge such uncertainty, but the intervals fail to account for it. Therefore, we look to extend the methods to a fully Bayesian analysis in Chapter 5.
In this chapter, we demonstrate how to extend the flexible GLMM specified in (4.11-4.13) in Chapter 4 to a fully Bayesian analysis, and how to carry out estimation and prediction in the Bayesian setting. Bayesian methods have been applied widely in random-effects meta-analysis. These methods provide natural advantages, especially in decision-making and in prediction. They also help resolve some of the existing problems, as reviewed in Section 2.5. We demonstrate the fully Bayesian approach with the Aprotinin dataset.

5.1 A Fully Bayesian Hierarchical Model

A fully Bayesian analysis places prior distributions on the unknown parameters. One possible set of prior distributions for parameters of the flexible GLMM can be
established as:

\[
\begin{align*}
\alpha & \sim Uniform(a_{\alpha,0}, b_{\alpha,0}) \\
\mu & \sim Normal(\mu_0, \sigma_{\mu,0}^2) \\
\theta & \sim Normal(\theta_0, \sigma_{\theta,0}^2) \\
\omega^2 & \sim Inv. Gamma(a_{\omega,0}, b_{\omega,0}) \\
\tau^2 & \sim Inv. Gamma(a_{\tau,0}, b_{\tau,0})
\end{align*}
\]

\[
\rho/2 + 0.5 \sim Beta(a_{\rho,0}, b_{\rho,0}).
\]

### 5.1.1 Hyper-Parameters

The Bayesian framework encourages the incorporation of prior information, and thus it makes sense to use more subjective priors that reflect the \textit{a priori} knowledge of the practitioners. We intend to choose the values of the hyper-parameters for the prior distributions from an “introspective” point of view, so the prior distributions “provide support to all reasonable parameters values” (Warn et al., 2002).

We initially specify the prior distribution for the parameter \( \alpha \) as a uniform distribution between 0 and 1, indicating that we do not have any prior belief about where \( \alpha \) lies. The priors on the two mean parameters (\( \mu \) and \( \theta \)) are both diffuse normal distributions centered at 0 with a standard deviation of 3. By centering \( \theta(\alpha) \) at zero, we take a conservative view towards the treatment being effective, and a standard deviation of 3 indicates that there is very small probability that the overall relative effect \( \theta(\alpha) \) is greater than 9 or smaller than -9 (3 standard deviations from the mean). A variance of 9 is huge on the logarithm scale. The prior on \( \mu \) can be translated as meaning that we believe \( \pi_c \) is centered at 0.5 and spreads widely across 0 and 1.
It has been discussed in literature that a ‘non-informative’ improper prior (e.g. $p(\tau^2) \propto 1/\tau^2$ or $IG(0, 0)$) for the variance parameters is not appropriate, especially for a small number of studies, since it may lead to an inappropriate posterior (Smith et al., 1995; Higgins and Spiegelhalter, 2002). For informative priors, Smith et al. (1995) proposed using an inverse gamma distribution $IG(1, 3)$ on the log odds scale; Gelman (2006) suggested the use of the weakly subjective half-t family; a locally uniform distribution on the standard deviation $\tau$ is also used (Warn et al., 2002). For this set of priors, we chose the inverse gamma distributions ($\alpha = 3$, $\beta = 4$) for both variance parameters $\tau^2$ and $\omega^2$. This distribution is flatter than the one used by Smith et al. (1995), which easily allows for reasonably large variability and is less subjective. Comparing the performances of different priors and making recommendations on prior specification is beyond the scope of this chapter, but it is certainly an interesting topic for further investigation.

A perfect correlation between $\mu_i(\alpha)$ and $\theta_i(\alpha)$ is hard to believe. Instead of giving $\rho$ a uniform prior between $-1$ and $1$, we put more mass near $\rho = 0$ by giving the transformation $\rho/2 + 0.5$ a Beta(3, 5) distribution.

Ideally, information from sources other than, but related to, the trials under consideration is available, which can be used in the choice of prior distributions (Higgins and Spiegelhalter, 2002). Alternatively, the priors can reflect the subjective clinical opinions of practitioners who truly understand the mechanism of the treatment and the data. For example, as discussed in Chapter 5, since the event in Aprotinin is ‘bleeding in cardiac surgery’, $\pi_t < \pi_c$ is preferred and it makes more sense to recruit higher risk patients (higher $\pi_c$). If we make an assumption that the control group event rate is 95% likely to be within the range 0.5 to 1, an informative distribution
of \( \mu \) reflecting such an \textit{a priori} belief is parametrized by \( \mu_0 = 1.1 \) and \( \sigma_{\mu,0} = 0.55 \) (at \( \alpha = 0.5 \)).

We present the results analyzing the Aprotinin dataset in Section 5.2 under the two sets of priors described above. In the first set (referred to as “priors 1” in the following discussion), the control group information \( \mu \) follows a normal distribution centered at 0 with a standard deviation equal to 3. The second set of priors (referred to as “priors 2” in the following discussion) gives \( \mu \) a more subjective prior distribution \( \mu \sim N(\mu_0 = 1.1, \sigma_{\mu,0} = 0.55) \), reflecting the fact that patients recruited for the Aprotinin studies should have larger risk of bleeding (high event rate in the control group). Prior distributions on the other parameters are the same.

\subsection*{5.1.2 Implementation and Inference}

Due to the complex hierarchical structure of the fully Bayesian model, we use the Metropolis-Hastings algorithm to obtain approximate samples from the posterior distribution \( p(\theta | y) \). For each parameter, we have a symmetric proposal distribution that is centered at the value from the previous iteration. For example, the proposal distribution for \( \mu^{(t)} \) is a uniform distribution with mean \( \mu^{(t-1)} \). Since the proposal distribution is symmetric, the acceptance ratio \( r \) is calculated by

\[
r = \frac{p(\mu^\ast, \theta | y) p(\theta^{(t-1)} | y)}{p(\mu^{(t-1)} | y) p(\theta | y)},
\]

where \( \mu^\ast \) is the proposed value for parameter \( \mu \) and \( \theta^{(t-1)} \) is the values of all parameters in the \((t-1)\text{th}\) iteration. \( \mu^\ast \) is accepted with probability \( min(1, r) \). The same sampling algorithm is applied for all of the parameters individually within each iteration. The acceptance rates of the proposed values have a range from 0.15 to 0.45.
Following a burn-in of 500,000 iterations, another 500,000 iterations were used to estimate the posterior quantities. The posterior samples of $\pi_t$ and $\pi_c$ are converted from $\theta$ and $\mu$ at each iteration. Credible intervals of each parameter are formed by taking the 2.5% and 97.5% quantiles of the corresponding MCMC samples as endpoints of the intervals.

Compared to the classical methods, Bayesian methods have a natural advantage in prediction. In the classical approach, we ignored the fact that $\alpha$ and the variance components are estimated from the data. The Bayesian approach, on the other hand, allows for uncertainty in all parameters in the model. A predictive sample for $\theta(\alpha)_{new}$ in a new study can be obtained by sampling from the conditional distribution

$$\theta(\alpha)_{new}|\mu(\alpha)^* \sim N\left(\theta + \frac{\tau}{\omega} \rho(\mu(\alpha)^* - \mu), \tau^2(1 - \rho^2)\right)$$  \hspace{1cm} (5.1)

at each iteration of the MCMC chain, where $\mu(\alpha)^* = g(\alpha, \pi_{c}^*)$, which results in a sample from the marginal posterior distributions of $\theta$. The 95% prediction interval is formed by taking the 2.5% and 97.5% quantiles of the simulated sample of $\theta(\alpha)_{new}$.

To find the prediction on the $\pi_t$ scale, we convert $\theta(\alpha)_{new}$ to $\pi_{t, new}$ at corresponding values of $\alpha$, where

$$\pi_{t, new} = g^{-1}(\alpha, \theta(\alpha) + \mu(\alpha)).$$  \hspace{1cm} (5.2)

5.2 Results

We use the 64 studies on the efficacy of Aprotinin to illustrate the fully Bayesian model constructed in Section 5.1.

We see in Table 5.1 that under both prior assumptions, the treatment group event rate $\pi_t$ is significantly smaller than $\pi_c$, which means that Aprotinin is effective in reducing bleeding during cardiac surgeries. Under the subjective prior assumption on $\mu$,
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Method</th>
<th>Mean</th>
<th>SD</th>
<th>95% credible interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>priors 1</td>
<td>0.47</td>
<td>0.10</td>
<td>(0.27,0.65)</td>
</tr>
<tr>
<td></td>
<td>priors 2</td>
<td>0.45</td>
<td>0.09</td>
<td>(0.27,0.62)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>priors 1</td>
<td>-0.63</td>
<td>0.09</td>
<td>(-0.80,-0.46)</td>
</tr>
<tr>
<td></td>
<td>priors 2</td>
<td>-0.65</td>
<td>0.09</td>
<td>(-0.82,-0.48)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>priors 1</td>
<td>0.53</td>
<td>0.23</td>
<td>(0.10,1.00)</td>
</tr>
<tr>
<td></td>
<td>priors 2</td>
<td>0.61</td>
<td>0.21</td>
<td>(0.20,1.04)</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>priors 1</td>
<td>0.32</td>
<td>0.07</td>
<td>(0.21,0.48)</td>
</tr>
<tr>
<td></td>
<td>priors 2</td>
<td>0.33</td>
<td>0.07</td>
<td>(0.21,0.59)</td>
</tr>
<tr>
<td>$\omega^2$</td>
<td>priors 1</td>
<td>1.14</td>
<td>0.52</td>
<td>(0.47,2.50)</td>
</tr>
<tr>
<td></td>
<td>priors 2</td>
<td>1.17</td>
<td>0.52</td>
<td>(0.49,2.49)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>priors 1</td>
<td>-0.66</td>
<td>0.12</td>
<td>(-0.83,-0.35)</td>
</tr>
<tr>
<td></td>
<td>priors 2</td>
<td>-0.67</td>
<td>0.11</td>
<td>(-0.83,-0.39)</td>
</tr>
<tr>
<td>$\pi_t$</td>
<td>priors 1</td>
<td>0.43</td>
<td>0.05</td>
<td>(0.34,0.53)</td>
</tr>
<tr>
<td></td>
<td>priors 2</td>
<td>0.44</td>
<td>0.05</td>
<td>(0.35,0.54)</td>
</tr>
<tr>
<td>$\pi_c$</td>
<td>priors 1</td>
<td>0.72</td>
<td>0.05</td>
<td>(0.61,0.82)</td>
</tr>
<tr>
<td></td>
<td>priors 2</td>
<td>0.74</td>
<td>0.05</td>
<td>(0.63,0.82)</td>
</tr>
</tbody>
</table>

Table 5.1: Results from using the Metropolis-Hastings method for fitting a fully Bayesian meta-analysis using the flexible GLMM for the Aprotinin dataset. Prior 1 is the initial prior distributions; $\mu$ has a more information prior distribution in prior 2.

the estimate of $\mu$ (0.61) and $\pi_c$ (0.74) are pulled to the right from the estimates under the first prior assumptions ($\mu = 0.53$, $\pi_c = 0.72$), but not by much. The estimates of the other parameters are also not largely affected by the subjective prior on $\mu$. This reflects that when we have enough data (64 studies in this case), the likelihood dominates the posterior distribution so any reasonable priors will give valid results.

With a more extreme (unreasonable) prior on $\mu$ ($\mu \sim N(\mu_0 = 2.5, \sigma_{\mu,0} = 0.7)$), the estimate of $\alpha$ was even smaller (0.39) compared to that in prior 2 (0.45). This more extreme prior for $\mu$ results in large influences on the estimates of the other parameters as well, but it is not reasonable as it forces $\pi_c$ to be mostly between 0.9 and 1. In
the future, we are interested in further exploring the influence of the priors of $\mu$, and other parameters, on the estimates of $\alpha$.

The predictions and prediction band for $\pi_t$ under various possible values of $\pi_c$ are presented in Figure 5.1, built from the results under priors 1 using (5.1) and (5.2). When $\pi_c$ is greater than 0.89, we predict $\pi_t$ to be strictly smaller than $\pi_c$ (the upper bound of the prediction interval of $\pi_t$ is smaller than $\pi_c$). When $\pi_c$ is between 0.14 and 0.89, we predict $\pi_t$ to be smaller than $\pi_c$, but the upper bounds of the prediction intervals of $\pi_t$ are larger than $\pi_c$. Aprotinin is predicted to be ineffective for patients whose risk of bleeding is lower than 0.14. The same conclusions are drawn from the predictions built from the results under priors 2.

Unlike the prediction curve in the classical model prediction results (Figure 4.12) which falls strictly below the diagonal line of no effect, the fully Bayesian model prediction curve is above the diagonal line when $\pi_c$ is small (less than 0.14). This is consistent with the dataset, in which both of the trials that have an observed control group event rate less than 0.14 have a higher observed treatment group event rate than the control group event rate.

The prediction band shown in Figure 5.1 from the fully Bayesian model is much wider than that from the classical model shown in Figure 4.12. In the classical flexible GLMM, we built the predictive distribution of $\theta_{\text{new}}(\alpha)$ for a fixed $\alpha = \hat{\alpha}$. The variance components $\tau$, $\omega$, and $\rho$ (estimated from the data) are assumed to be fixed as well. The uncertainties in estimating these parameters are not accounted for. The Bayesian prediction procedure accounts for all the uncertainties, which results in much wider prediction intervals.
Figure 5.1: Prediction curve of $\pi_t$ given values of $\pi_c$ using results from the fully Bayesian model (priors 1). The plots on the side and on the bottom show the posterior distributions of the medians of $\pi_t$ and $\pi_c$. 
For further investigation into whether the uncertainty in $\alpha$ is a major source of the overall uncertainty, we fit the Bayesian model using prior 1 with $\alpha$ fixed at $\hat{\alpha} = 0.595$, the estimated $\alpha$ under the classical model. When we give $\alpha$ a uniform prior $\alpha \sim \text{unif}(0, 1)$, the estimates or predictions are a weighted average of the results across all values of $\alpha$. We see from Figure 5.2 that fixing $\alpha$ has a great influence on the prediction curve and the prediction curve is less conservative. Further investigation on the impact of the prior distribution of $\alpha$ is left for future work.

5.3 Summary

This chapter extends the flexible GLMM to a Bayesian structure. Using MCMC algorithms, the fully Bayesian analysis avoids the need for approximating the likelihood function. It adds more flexibility to the (already) flexible modeling structure we proposed in Chapter 4. In the classical approach, the uncertainty in estimating $\alpha$ and the variance components are ignored as they were considered fixed in building the predictive distributions. The Bayesian model, on the other hand, allows for uncertainty in all parameters in the model, as well as the uncertainty in choosing a model. Credible intervals for the parameters are obtained from the posterior distributions and building predictive distributions in Bayesian approaches is straightforward through simulation.

The choice of priors is always a topic under debate for Bayesian analysis. Recognizing that non-informative priors may not be appropriate for a meta-analysis, we used two sets of informative priors. The first set of priors reflects conservative and moderate $a priori$ opinions, while the second set reflects a stronger opinion on the control group event rate by considering a context-specific reasonable range. Under
Figure 5.2: Prediction curve of $\pi_t$ given values of $\pi_c$ using results from the fully Bayesian model (priors 1) except that $\alpha$ is fixed at 0.595. The plots on the side and on the bottom show the posterior distributions of the medians of $\pi_t$ and $\pi_c$. 
both sets of priors, Aprotinin is estimated to be effective and the choice of prior did not influence the estimation and prediction results much. Further investigations on the choice of prior distributions should be carried when using this model in practice.
Chapter 6: Contributions and Future Work

Meta-analysis is a statistical method used to quantitatively synthesize multiple independent studies that address a same or similar problem. Moment-based estimators combining a summary statistic from each individual study are widely used due to their simplicity. There are two assumptions associated with the moment-based methods: fixed- and random-effect assumptions. The fixed-effect methods assume that all studies share a common treatment effect, while the random-effect methods allow the true effect of the individual studies to vary. There has been a long debate between the two assumptions, but in general, the random-effect assumption is considered to be more realistic. The moment-based estimators have been shown to be biased, especially in estimating the heterogeneity parameter $\tau^2$. For studies with binary outcomes (yes/no), the bias is even more prominent when the events are rare. As some of the common relative measures (e.g. odds ratio and relative risk) are not defined when one or both arms of a study have zero events, such studies are either excluded from a moment-based meta-analysis or included with an arbitrary continuity correction.

The likelihood-based methods that are used to model the binomial structure of the data have various advantages over the moment-based methods in estimating an overall treatment effect, testing for heterogeneity and predicting a future study. However,
the likelihood-based method requires model specification. This dissertation examines current and proposed new methods for hierarchical model selection.

In Chapter 3, we compare conditional models to joint models. A conditional model is often used for the reason that some nuisance parameters can be eliminated when conditioning on the total number of events in a study. One conditional approach is a non-central hypergeometric normal model proposed by Sidik and Jonkman (2008). We compare the SJ conditional model to a full conditional approach derived from a joint hierarchical bivariate normal model and conclude that the SJ conditioning approach is not a special case of the full conditioning approach. Conditioning in the joint model does not eliminate the nuisance parameters or simplify the computation as it does in the SJ model. We recommend that the use of conditioning as a parameter reducing approach should be justified by more than the mathematical convenience. A random-effect joint model that captures the binomial structure of the data and the dependence relationship between the treatment and the control group should be the approach taken.

In Chapter 4, we extend the joint model with correlated random treatment effect and random intercept (control group log odds) to a flexible model sharing the same structure. The flexible model is constructed based on a newly proposed summary scale \( \theta(\alpha) \), a transformation of the original scale of \( \pi_t \) and \( \pi_c \) with an unknown parameter \( \alpha \). For a meta-analysis with a moderate number of studies, we estimate \( \alpha \) to find the scale that best fits the observed data. A general framework based on the probability distribution of the observed data across the two-dimensional space of \( \pi_{ti} \) and \( \pi_{ci} \) is described. A simple and intuitive prediction approach on the scale of \( \pi_t \) and \( \pi_c \) is proposed. Predictions on this scale can be transformed to any scales of interest.
The flexible modeling framework proposed in Chapter 4 inherits the advantages of the likelihood-based joint models, and acknowledges the uncertainty in model specification. We extend it to a fully Bayesian analysis in Chapter 5, so it has the additional flexibility of incorporating full uncertainty in all parameters. The Bayesian analysis is also advantageous for meta-analyses with very small numbers of studies. It is particularly hard to estimate the between-study variance and to estimate $\alpha$. Using external information on reasonable values for $\tau^2$ and $\alpha$ may be more desirable than estimating them from a limited amount of data. The Bayesian structure allows such external information or a priori beliefs.

Beyond this dissertation, there are many very interesting extensions and problems for further investigation:

- The proposed model can incorporate additional covariates. The current model has a study-level covariate indicating the baseline risk. Another covariate that may be influential is the size of the study. When individual patient data (IPD) is available, the individual-level covariates can be included as well. A large collection of studies will be required to evaluate the impact of the covariates.

- In Chapter 4, we only report a few limited simulations to compare the flexible GLMM model to the traditional one, and to validate the former model’s implementation. More extensive studies can be carried out to investigate the type I error and the power for the new model.

- Studying the mechanism behind the choice of $\alpha$ will be interesting. We see from the four examples in Chapter 4 that $\alpha$ tends to be larger when the observed control group event rates are higher (usually associated with harmful outcomes).
The impact of the prior distribution of $\pi_c$ on the estimation of $\alpha$ is also observed in Chapter 5.

- In Chapter 5, we extended the flexible GLMM model to a fully Bayesian analysis and demonstrated how estimation and prediction can be done in a Bayesian setting. We compared two sets of relatively subjective prior distributions that appear to give reasonable results. Further work could compare different prior choices, especially for the variance parameters, carry out a sensitivity analysis and make recommendations for prior specification.

- The flexible framework we proposed only incorporated the relative scales (the risk ratio, the odds ratio and the reversed risk ratio). When events are rare, the three measures do not differ by much. Developing an even more flexible framework that incorporates the absolute measure is attractive for meta-analysis for sparse events in safety evaluation.

- The two-dimensional framework can also be used for the estimation and prediction of other interesting quantities, for example, an average across a mixture of future studies.

Meta-analysis has become an important statistical tool in medical practice and drug development for both efficacy and safety evaluation. Many practitioners, however, choose convenient estimators that are easy to implement and pay less or no attention to the model selection and particularly, the interpretation of the results.

It has been commonly agreed in literature that the convenient moment-based estimators are less than ideal, especially when the studies are heterogeneous and when events are rare. Alternative approaches include likelihood-based methods, for
which model specification is key. Depending on how the model is specified, there may be differences in how the results are interpreted.

We investigate current methods and proposed new methods for random-effect hierarchical model selection. A general framework is set up to think about the probability distributions of the data across studies on the scale of $\pi_{ti}$ (the study-specific treatment event rate) and $\pi_{ci}$ (the study-specific control event rate). Interpretation on this original scale is straightforward and results can be transformed into any scale for further inference.
Appendix A: The Endoscopic Hemostasis Dataset

The dataset contains a collection of 41 randomized trials on a new surgical therapy for bleeding peptic ulcers conducted between 1980 and 1989 (Sacks et al., 1990; Efron, 1996; Sidik and Jonkman, 2008). Studies 40 and 41 are excluded from the analysis in Chapter 3. All 41 studies are used in the analysis in Chapter 4.

<table>
<thead>
<tr>
<th>Study</th>
<th>Treatment No. of Events</th>
<th>Treatment No. of Patients</th>
<th>Control No. of Events</th>
<th>Control No. of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>8</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>11</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>29</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>29</td>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>13</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>15</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>11</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>36</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>12</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>14</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>22</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>30</td>
<td>4</td>
<td>23</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>24</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>36</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td>21</td>
<td>6</td>
<td>34</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>22</td>
<td>4</td>
<td>14</td>
<td>5</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>23</td>
<td>14</td>
<td>54</td>
<td>13</td>
<td>61</td>
</tr>
<tr>
<td>24</td>
<td>6</td>
<td>15</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>9</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>27</td>
<td>5</td>
<td>12</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>10</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>29</td>
<td>0</td>
<td>22</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>16</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>14</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>32</td>
<td>8</td>
<td>16</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>33</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>34</td>
<td>0</td>
<td>20</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>35</td>
<td>4</td>
<td>13</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>36</td>
<td>10</td>
<td>30</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>37</td>
<td>3</td>
<td>13</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>38</td>
<td>4</td>
<td>30</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>39</td>
<td>7</td>
<td>31</td>
<td>15</td>
<td>22</td>
</tr>
<tr>
<td>40</td>
<td>0</td>
<td>34</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>
Appendix B: The Aprotinin Dataset

The dataset contains a collection of 64 randomized trials on Aprotinin from 1987 to 2002 with an endpoint of perioperative transfusion (Fergusson et al., 2005). Aprotinin is an inhibitor used to reduce perioperative bleeding and the need for blood transfusion during cardiac surgeries. The trials included in this dataset show great heterogeneity: the range of the trial sizes is from 20 to 1784, and the range of the observed control group event rates is from 0 to 1.

Fergusson et al. (2005) used this dataset to demonstrate that the researchers studying the efficacy of Aprotinin did not take advantage of the previous published evidence. They carried out a cumulative meta-analysis and argued that the efficacy of Aprotinin was definitely address by the first 12 studies and the following RCTs addressing the same question were not necessary.

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No. of Events</td>
<td>No. of Patients</td>
</tr>
<tr>
<td>1</td>
<td>1987</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>1989</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>1989</td>
<td>16</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>1990</td>
<td>15</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>1990</td>
<td>11</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>1990</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>7</td>
<td>1991</td>
<td>29</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>1991</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>1991</td>
<td>8</td>
<td>30</td>
</tr>
<tr>
<td>Year</td>
<td>Year Result</td>
<td>Result A</td>
<td>Result B</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>1991</td>
<td>10</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>1992</td>
<td>11</td>
<td>35</td>
<td>58</td>
</tr>
<tr>
<td>1992</td>
<td>12</td>
<td>549</td>
<td>902</td>
</tr>
<tr>
<td>1992</td>
<td>13</td>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>1992</td>
<td>14</td>
<td>32</td>
<td>34</td>
</tr>
<tr>
<td>1992</td>
<td>15</td>
<td>55</td>
<td>113</td>
</tr>
<tr>
<td>1993</td>
<td>16</td>
<td>9</td>
<td>43</td>
</tr>
<tr>
<td>1993</td>
<td>17</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>1993</td>
<td>18</td>
<td>55</td>
<td>140</td>
</tr>
<tr>
<td>1993</td>
<td>19</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>1994</td>
<td>20</td>
<td>42</td>
<td>51</td>
</tr>
<tr>
<td>1994</td>
<td>21</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>1994</td>
<td>22</td>
<td>28</td>
<td>74</td>
</tr>
<tr>
<td>1994</td>
<td>23</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>1994</td>
<td>24</td>
<td>51</td>
<td>54</td>
</tr>
<tr>
<td>1994</td>
<td>25</td>
<td>68</td>
<td>75</td>
</tr>
<tr>
<td>1994</td>
<td>26</td>
<td>5</td>
<td>61</td>
</tr>
<tr>
<td>1994</td>
<td>27</td>
<td>22</td>
<td>28</td>
</tr>
<tr>
<td>1994</td>
<td>28</td>
<td>33</td>
<td>49</td>
</tr>
<tr>
<td>1994</td>
<td>29</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>1994</td>
<td>30</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>1994</td>
<td>31</td>
<td>60</td>
<td>110</td>
</tr>
<tr>
<td>1995</td>
<td>32</td>
<td>7</td>
<td>49</td>
</tr>
<tr>
<td>1995</td>
<td>33</td>
<td>15</td>
<td>43</td>
</tr>
<tr>
<td>1995</td>
<td>34</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>1995</td>
<td>35</td>
<td>25</td>
<td>51</td>
</tr>
<tr>
<td>1995</td>
<td>36</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>1995</td>
<td>37</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>1995</td>
<td>38</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>1995</td>
<td>39</td>
<td>12</td>
<td>47</td>
</tr>
<tr>
<td>1995</td>
<td>40</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>1995</td>
<td>41</td>
<td>109</td>
<td>189</td>
</tr>
<tr>
<td>1996</td>
<td>42</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>1996</td>
<td>43</td>
<td>43</td>
<td>46</td>
</tr>
<tr>
<td>1996</td>
<td>44</td>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td>1996</td>
<td>45</td>
<td>55</td>
<td>75</td>
</tr>
<tr>
<td>1996</td>
<td>46</td>
<td>73</td>
<td>127</td>
</tr>
<tr>
<td>1996</td>
<td>47</td>
<td>45</td>
<td>487</td>
</tr>
<tr>
<td>1997</td>
<td>48</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>1997</td>
<td>49</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>1997</td>
<td>50</td>
<td>66</td>
<td>110</td>
</tr>
<tr>
<td>1997</td>
<td>51</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>1997</td>
<td>52</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>No.</td>
<td>Year</td>
<td>Value1</td>
<td>Value2</td>
</tr>
<tr>
<td>-----</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>53</td>
<td>1998</td>
<td>10</td>
<td>37</td>
</tr>
<tr>
<td>54</td>
<td>1998</td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>55</td>
<td>1998</td>
<td>152</td>
<td>401</td>
</tr>
<tr>
<td>56</td>
<td>1999</td>
<td>23</td>
<td>36</td>
</tr>
<tr>
<td>57</td>
<td>1999</td>
<td>24</td>
<td>100</td>
</tr>
<tr>
<td>58</td>
<td>2000</td>
<td>6</td>
<td>40</td>
</tr>
<tr>
<td>59</td>
<td>2000</td>
<td>6</td>
<td>56</td>
</tr>
<tr>
<td>60</td>
<td>2000</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>61</td>
<td>2001</td>
<td>37</td>
<td>101</td>
</tr>
<tr>
<td>62</td>
<td>2001</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>63</td>
<td>2001</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>64</td>
<td>2002</td>
<td>2</td>
<td>15</td>
</tr>
</tbody>
</table>
Appendix C: Consistency of Different Measures (Two Examples)

We present two example meta-analyses to illustrate the study-to-study variation related to different measures of treatment effect. The datasets used are the Fergusson et al. (2005) study of Aprotinin (section 1.2.4) and the meta-analysis of Lamotrigine add-on therapy for treatment of epilepsy (section 1.2.4). The second example is a reproduction of Example 4 in Deeks (2002). The different summary metrics of odds ratio (OR), risk ratio (RR), reversed risk ratio (RR.r) and risk difference (RD) are estimated using both the MH method (fixed-effect) and the DL method (random-effect). Cochran’s $\chi^2$ test is used to test the null hypothesis of homogeneity for each summary metric. The results are shown in Table C.1.

We see from the results that Aprotinin significantly reduces the risk of perioperative bleeding during cardiac surgeries under both models for all four measures of effect. Heterogeneity presents in all four measures (rejection of the null hypothesis of no heterogeneity). Although the Q-statistics for the four measures are all big enough to reject the null hypothesis, they differ significantly. OR gives the smallest value ($Q = 115.58$) and RR gives the biggest one ($Q = 454.86$).

Lamotrigine add-on is also concluded to be significantly beneficial using all four measures under both models. Heterogeneity only presents in the measure RR.r, the
<table>
<thead>
<tr>
<th>Measure</th>
<th>Model</th>
<th>Effect</th>
<th>95% CI</th>
<th>Heterogeneity test</th>
<th>Effect</th>
<th>95% CI</th>
<th>Heterogeneity test</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR</td>
<td>Fixed</td>
<td>0.36</td>
<td>[0.32, 0.40]</td>
<td>$Q = 115.58$</td>
<td>2.87</td>
<td>[1.92, 4.29]</td>
<td>$Q = 5.71$</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>0.34</td>
<td>[0.29, 0.41]</td>
<td>$d.f. = 62$</td>
<td>2.83</td>
<td>[1.89, 4.25]</td>
<td>$d.f. = 9$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p-value &lt; 0.0001</td>
<td></td>
<td></td>
<td>p-value = 0.77</td>
</tr>
<tr>
<td>RR</td>
<td>Fixed</td>
<td>0.70</td>
<td>[0.67, 0.72]</td>
<td>$Q = 454.86$</td>
<td>2.33</td>
<td>[1.67, 3.23]</td>
<td>$Q = 4.62$</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>0.69</td>
<td>[0.63, 0.75]</td>
<td>$d.f. = 63$</td>
<td>2.28</td>
<td>[1.63, 3.18]</td>
<td>$d.f. = 9$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p-value &lt; 0.0001</td>
<td></td>
<td></td>
<td>p-value = 0.87</td>
</tr>
<tr>
<td>RR.r</td>
<td>Fixed</td>
<td>1.59</td>
<td>[1.51, 1.67]</td>
<td>$Q = 290.58$</td>
<td>0.84</td>
<td>[0.78, 0.89]</td>
<td>$Q = 17.24$</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>1.59</td>
<td>[1.42, 1.78]</td>
<td>$d.f. = 62$</td>
<td>0.87</td>
<td>[0.80, 0.95]</td>
<td>$d.f. = 9$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p-value &lt; 0.0001</td>
<td></td>
<td></td>
<td>p-value = 0.045</td>
</tr>
<tr>
<td>RD</td>
<td>Fixed</td>
<td>-0.20</td>
<td>[-0.22, -0.18]</td>
<td>$Q = 258.34$</td>
<td>0.15</td>
<td>[0.10, 0.20]</td>
<td>$Q = 14.72$</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>-0.20</td>
<td>[-0.24, -0.16]</td>
<td>$d.f. = 63$</td>
<td>0.12</td>
<td>[0.06, 0.19]</td>
<td>$d.f. = 9$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p-value &lt; 0.0001</td>
<td></td>
<td></td>
<td>p-value = 0.10</td>
</tr>
</tbody>
</table>

Table C.1: Evaluation of the heterogeneity in different summary metrics using Cochran’s $\chi^2$ test and the Q-statistic.
reverse risk ratio. The p-values of OR and RR are large, indicating that the two measures are relatively consistent across studies.
Appendix D: Empirical Evidence of the Dependency between the Summary Statistics and the Empirical Weights

To explore the dependence between the summary statistics and the empirical weights, we use the log odds ratio as an example. For the DL effect estimator, the study-specific weight used is

$$w_i = \hat{\tau}^2 + \hat{\sigma}_{i}^2,$$

where the between-study variance estimator $\hat{\tau}^2$ is a function of $\hat{\sigma}_{i}^2$. Therefore, the problem can be simplified to studying the correlation between the observed odds ratio of study $i$

$$\hat{\theta}_i = \log \frac{r_{ti}/(n_{ti} - r_{ti})}{r_{ci}/(n_{ci} - r_{ci})}$$

and the reciprocal of its sampling variance

$$\hat{\sigma}_{i}^{-2} = \left( \frac{1}{r_{ti}} + \frac{1}{(n_{ti} - r_{ti})} + \frac{1}{r_{ci}} + \frac{1}{(n_{ci} - r_{ci})} \right)^{-1}.$$

Assuming

$$r_{ti} \sim Bin(n_{ti}, \pi_{ti}) \quad (D.1)$$

$$r_{ci} \sim Bin(n_{ci}, \pi_{ci}),$$

$Cov(\hat{\theta}_i, \hat{\sigma}_{i}^{-2})$ can be found mathematically using a large-sample approximation and the delta method. The observed numbers of events for a single study are simulated from model (D.1) using different values of $\pi_{ti}$ and $\pi_{ci}$. The number of patients in each
Table D.1: Empirical correlation coefficients of the observed log odds ratio and the reciprocal of its sampling variance (one study).

<table>
<thead>
<tr>
<th>$\pi_t$</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>-0.003</td>
<td>0.588</td>
<td>0.764</td>
<td>0.833</td>
<td>0.869</td>
<td>0.892</td>
<td>0.914</td>
<td>0.947</td>
<td>0.983</td>
</tr>
<tr>
<td>0.2</td>
<td>-0.590</td>
<td>-0.001</td>
<td>0.424</td>
<td>0.653</td>
<td>0.784</td>
<td>0.875</td>
<td>0.951</td>
<td>0.991</td>
<td>0.947</td>
</tr>
<tr>
<td>0.3</td>
<td>-0.764</td>
<td>-0.425</td>
<td>0.001</td>
<td>0.412</td>
<td>0.725</td>
<td>0.918</td>
<td>0.986</td>
<td>0.951</td>
<td>0.914</td>
</tr>
<tr>
<td>0.4</td>
<td>-0.833</td>
<td>-0.652</td>
<td>-0.411</td>
<td>0.001</td>
<td>0.645</td>
<td>0.945</td>
<td>0.917</td>
<td>0.875</td>
<td>0.893</td>
</tr>
<tr>
<td>0.5</td>
<td>-0.869</td>
<td>-0.783</td>
<td>-0.724</td>
<td>-0.646</td>
<td>0.003</td>
<td>0.645</td>
<td>0.724</td>
<td>0.784</td>
<td>0.868</td>
</tr>
<tr>
<td>0.6</td>
<td>-0.893</td>
<td>-0.875</td>
<td>-0.918</td>
<td>-0.945</td>
<td>-0.645</td>
<td>0.001</td>
<td>0.411</td>
<td>0.652</td>
<td>0.833</td>
</tr>
<tr>
<td>0.7</td>
<td>-0.914</td>
<td>-0.951</td>
<td>-0.986</td>
<td>-0.917</td>
<td>-0.725</td>
<td>-0.411</td>
<td>0.000</td>
<td>0.424</td>
<td>0.765</td>
</tr>
<tr>
<td>0.8</td>
<td>-0.947</td>
<td>-0.991</td>
<td>-0.951</td>
<td>-0.875</td>
<td>-0.784</td>
<td>-0.653</td>
<td>-0.422</td>
<td>0.001</td>
<td>0.588</td>
</tr>
<tr>
<td>0.9</td>
<td>-0.983</td>
<td>-0.947</td>
<td>-0.914</td>
<td>-0.892</td>
<td>-0.868</td>
<td>-0.833</td>
<td>-0.763</td>
<td>-0.590</td>
<td>0.000</td>
</tr>
</tbody>
</table>

$\pi_t$ is the true event rate in treatment group and $\pi_c$ is the true event rate in the control group.

group are set to be equal ($n_{ti} = n_{ci} = 100$). The empirical correlations for different combinations of true event rates are shown in Table D.1.

For an individual study, the summary statistic $\hat{\theta}_i$ and the study-specific weight $w_i$ are highly correlated except on the diagonal of Table D.1, where the treatment group and the control group have the same event rate ($\pi_t = \pi_c$). In other words, they are highly correlated except when there is no treatment effect present, corresponding to the origin (0, 0) on the plot in Figure D.1. Figure D.1 shows the relationship of the true log odds ratio (treatment effect) and the empirical correlation. One can see that the correlations are positive when the true effect is less than 0 and negative when the true effect is greater than 0. This indicates that the individual treatment effects with bigger magnitudes are assigned larger weights, possibly resulting in a positive bias in the treatment-effect estimator.
Figure D.1: The relationship between the true log odds ratio and the empirical correlation coefficient.
Appendix E: A Simulation Study on the Performance of the Non-Central Hypergeometric Model

We conduct a simulation study to compare the performance of the non-central hypergeometric distribution in the estimation of the mean effect (log odds ratio, $\theta$) and the between-study variance ($\tau^2$) to the DL method and the binomial-normal random-effect model (3.6). The latter is also referred as the logistic model. In each simulation replication, $k = 100$ studies are simulated from model (3.6). The number of patients in the two arms of each study are set to be equal, generated by rounding a random draw from a normal distribution with mean 100 and standard deviation 5.

$\theta_i$ and $\phi_{ci}$ are sampled from a bivariate normal distribution for given values of $\theta$, $\phi_c$, $\tau^2$, $\omega^2$ and $\rho$. In the simulation we carry out, $\phi_c$, $\tau^2$ and $\omega^2$ are set to be fixed at $\phi_c = \log_{10}^{1.9}$, $\tau^2 = 1$, and $\omega^2 = 0.5$. We explore the influence of the magnitude and direction of $\theta$ (the true overall effect) and $\rho$ (the correlation coefficient) on the estimates of $\theta$ and $\tau^2$. Values of $\theta$ and $\rho$ are both chosen to be from largely negative to largely positive: $\theta = (-3, -1, 0, 1, 3)$ and $\rho = (-0.9, -0.5, 0, 0.5, 0.9)$. 2000 replicate analyses of 100 simulated studies are performed for each combination of parameter values.

Figure E.1a shows the mean effect estimates as a function of $\rho$ when there is no treatment effect present ($\theta = 0$). The logistic model performs very well (the estimates
are all close to the purple line of the true $\theta$), while the NHG model and the DL model both show apparent biases in the mean effect estimation. As $\rho$ decreases, the biases of the DL estimate and the NHG estimate go from largely positive to slightly negative. The biases produced by the NHG model are consistently smaller than those of the DL model. Assuming that the linear relationships hold, the line of the NHG model intersects with the purple line showing the true effect at roughly the same spot as the line connecting the DL estimates. This suggests that both the NHG and the DL mean effect estimator would be unbiased at $\rho \approx -0.7$, and that the NHG model may have the same bias mechanism related to $\rho$ as the DL method does.

![Mean effect estimates](image1)

**Figure E.1:** Mean effect estimates of the non-central hypergeometric model, the DL model and the logistic model. The purple dotted lines indicate the true effect $\theta = 0$ and $\hat{\theta} = \theta$ in the two plots, respectively. Estimates closer to the purple line are more desirable.
Although Figure E.1a clearly indicates that the NHG and DL estimates are biased (for $\theta = 0$) when $\theta_i$ and $\phi_{ci}$ are independent ($\rho = 0$), other independences are implied at the point of unbiasedness ($\rho = 0.7$.) In (3.11) and (3.12), we showed that $\rho = -\frac{\omega}{\tau}$ indicates the independence of the treatment log odds $\phi_t$ and control log odds $\phi_c$; and $\rho = -\frac{\tau}{2\omega}$ indicates the independence of the treatment effect $\theta_i$ and average rate $\frac{\phi_t + \phi_c}{2}$. In this simulation, we used $\tau^2 = 1$ and $\omega^2 = 0.5$, which gives

$$-\frac{\omega}{\tau} = -\frac{\tau}{2\omega} = \frac{1}{\sqrt{2}} \approx -0.7.$$  

This implies that the DL estimator is unbiased under at least one of the independence assumptions above. Unfortunately, with our set-up of $\tau^2 = 2\omega^2$, we are not able to tell which independence assumption gives the unbiased DL estimator. Further investigation distinguishing the two types of independence is left for future work.

Figure E.2: Bias of the mean effect estimates of the non-central hypergeometric model, the DL model and the logistic model at $\rho = 0$ for $\theta = (-3, -1, 0, 1, 3)$. The purple dotted line indicates no bias.
Figure E.1b further explores the case of independence between \( \theta_i \) and \( \phi_{ci} \). It shows the mean effect estimates as a function of the true effect \( \theta \) at \( \rho = 0 \). Estimates of the logistic model appears to be roughly on the diagonal purple line indicating equality of the estimates to the true value of \( \theta \). But, the scale of the plot obscures the bias of both NHG and DL estimates. To see the biases more clearly, we present them directly in Figure E.2. The DL estimate of \( \theta \) goes from largely upward biased to slightly downward biased as the true \( \theta \) increases from \(-3\) to \(3\). On the other hand, the biases in the NHG method consistently overestimate the true treatment effect when \( \rho \) is zero and the bias increases as \( \theta \) increases. Note that the pattern for NGH and DL are the same for all the other plots, but different in this case.

![Figures](a) \( \theta = 0 \) (b) \( \rho = 0 \)

Figure E.3: Between-study variance estimates of the non-central hypergeometric model, the DL model and the logistic model. The purple dotted line indicates the true variance \( \tau^2 = 1 \). Estimates closer to the purple line is more desirable.
The between-study variance ($\tau^2$) estimates are presented in Figure E.3. For both fixed $\theta = 0$ and $\rho = 0$, the DL model consistently underestimates the true $\tau^2 = 1$ (the purple dotted line). However, the NHG model somehow manages to capture the between-study variance correctly, especially when $\rho = 0$ and $\theta = 0$, as shown in Figure E.3b (although recall it is biased in estimating the overall mean effect). This suggests that the biasedness of the mean effect estimator should not be attributed to a model’s lack of ability to estimate the between-study variation.
Appendix F: The Approximate Distribution of the Flexible Measure

Assume that the numbers of events in each group of a study follow the binomial distributions:

\[ r_t | n_t \sim \text{Bin}(n_t, p_t) \]
\[ r_c | n_c \sim \text{Bin}(n_c, p_c). \]

Based on the large sample approximation and the delta method, we find

\[ \left( \frac{r_t}{n_t}, \frac{r_c}{n_c} \right) \sim \text{Normal} \left( \left( \frac{p_t}{p_c} \right), \left( \begin{array}{cc} \frac{p_t(1-p_t)}{n_t} & 0 \\ 0 & \frac{p_c(1-p_c)}{n_c} \end{array} \right) \right). \]

For a given \( \alpha \), \( \hat{\theta}(\alpha) \) is a function of the observe event rates \( r_t/n_t \) and \( r_c/n_c \):

\[
g \left( \frac{r_t}{n_t}, \frac{r_c}{n_c} \right) = \alpha \log \frac{r_t/n_t}{r_c/n_c} - (1 - \alpha) \log \frac{1 - r_t/n_t}{1 - r_c/n_c} = \hat{\theta}(\alpha).\]

To apply the delta method, letting

\[
g(a, b) = \alpha \log \frac{a}{b} - (1 - \alpha) \log \frac{1 - a}{1 - b},\]

we can find the mean of \( \hat{\theta}(\alpha) \) by

\[
\mu = g(p_t, p_c) = \alpha \log \frac{p_t}{p_c} - (1 - \alpha) \log \frac{1 - p_t}{1 - p_c},\]

127
and the variance by the following derivation:

\[
\frac{\partial g(a,b)}{\partial a} = \alpha b \frac{1}{a b} - (1 - \alpha) \frac{1}{1 - a} \frac{1}{1 - b} \frac{1}{-1} = \alpha \frac{1}{a} + (1 - \alpha) \frac{1}{1 - a}
\]

\[
\frac{\partial g(a,b)}{\partial b} = \alpha \frac{b}{a} \left(-\frac{a}{b^2}\right) - (1 - \alpha) \frac{1}{1 - a} \frac{1}{(1 - b)^2} = -\alpha \frac{1}{b} - (1 - \alpha) \frac{1}{1 - b}
\]

\[
\begin{pmatrix}
g'_a(p_t,p_c) \\
g'_b(p_t,p_c)
\end{pmatrix} = \begin{pmatrix}
\alpha \frac{1}{p_t} + (1 - \alpha) \frac{1}{1-p_t} \\
-\alpha \frac{1}{p_c} - (1 - \alpha) \frac{1}{1-p_c}
\end{pmatrix}
\begin{pmatrix}
\frac{p_t(1-p_t)}{n_t} \\
0
\end{pmatrix}
\begin{pmatrix}
\alpha \frac{1}{p_t} + (1 - \alpha) \frac{1}{1-p_t} \\
-\alpha \frac{1}{p_c} - (1 - \alpha) \frac{1}{1-p_c}
\end{pmatrix}
\begin{pmatrix}
\frac{p_t}{n_t(1-p_t)} + \frac{p_c}{n_c(1-p_c)}
\end{pmatrix}
\]

\[
\sigma^2 = \alpha^2 \left( \frac{1-p_t}{n_t p_t} + \frac{1-p_c}{n_c p_c} \right) + (1 - \alpha)^2 \left( \frac{p_t}{n_t (1-p_t)} + \frac{p_c}{n_c (1-p_c)} \right)
\]

\[
+ 2\alpha (1 - \alpha) \left( \frac{1}{n_t} + \frac{1}{n_c} \right).
\]

We can then write the approximate distribution of \( \hat{\theta}(\alpha) \) as

\[
\hat{\theta}(\alpha) = \alpha \hat{\eta} - (1 - \alpha) \hat{\eta}_R \sim Normal(\mu, \sigma^2),
\]

where

\[
\mu = \alpha \log \frac{p_t}{p_c} - (1 - \alpha) \log \frac{1-p_t}{1-p_c};
\]

\[
\sigma^2 = \alpha^2 \left( \frac{1-p_t}{n_t p_t} + \frac{1-p_c}{n_c p_c} \right) + (1 - \alpha)^2 \left( \frac{p_t}{n_t (1-p_t)} + \frac{p_c}{n_c (1-p_c)} \right)
\]

\[
+ 2\alpha (1 - \alpha) \left( \frac{1}{n_t} + \frac{1}{n_c} \right).
\]

128
The sampling mean and standard error are

\[ \hat{\mu} = \alpha \log \frac{r_t/n_t}{r_c/n_c} - (1 - \alpha) \log \frac{1 - r_t/n_t}{1 - r_c/n_c}, \]  
\[ \hat{\sigma}^2 = \alpha^2 \left( \frac{1}{r_t} - \frac{1}{n_t} + \frac{1}{r_c} - \frac{1}{n_c} \right) \]

\[ + (1 - \alpha)^2 \left( \frac{1}{n_t - n_t} - \frac{1}{n_t} + \frac{1}{n_c - r_c} - \frac{1}{n_c} \right) + 2\alpha(1 - \alpha) \left( \frac{1}{n_t} + \frac{1}{n_c} \right). \]
Appendix G: R code: User Specified Link Function using Package ‘lme4’

linkalpha <- function(alpha = 0.5)
{
    #link function
    linkfun <- function(p){
        alpha*log(p)-(1-alpha)*log(1-p)
    }

    # numerical evaluation to the inverse link function
    linkinv <- function(eta)
    {
        p.grid <- seq(10^(-10),1-10^(-10),length=10^5)
        g.p <- linkfun(p.grid)
        if(any(eta <= min(g.p))){
            trunc.low=1
            index.out.low <- which(eta < min(g.p))
            eta[index.out.low] <- min(g.p)
        }
        print("etahat is truncated on the left")
    }
}
if(any(eta >= max(g.p))){
    trunc.up <- 1
    index.out.up <- which(eta > max(g.p))
    eta[index.out.up] <- max(g.p)
    print("etahat is truncated on the right")
}

p.result <- vector(length=length(eta))

for(i in 1:length(eta)){
    if(any(g.p==eta[i])){
        p.result[i] <- p.grid[which(g.p==eta[i])]
    }else{
        ind.change <- max(which(g.p < eta[i]))
        p.result[i] <- mean(c(p.grid[ind.change],p.grid[ind.change+1]))
    }
}

return(p.result)

# numerical evaluation to the first derivative of the inverse link function
mu.eta <- function(eta, mat=inv.deri.grid.result)
{

deri.eta.result <- vector(length=length(eta))
eta.x <- mat[,1]
deri.eta.x <- mat[,2]
for (i in 1:length(eta)){
    if(eta[i] <= min(eta.x)){
        deri.eta.result[i] <- deri.eta.x[1]
    } else if(eta[i] >= max(eta.x)){
        deri.eta.result[i] <- deri.eta.x[length(eta.x)]
    } else{
        ind.low <- max(which(eta.x < eta[i]))
        #index.up <- min(which(eta.x > eta[i]))
        #interpolate
        deri.eta.result[i] <- deri.eta.x[ind.low]+(deri.eta.x[ind.low+1]
        -deri.eta.x[ind.low])*(eta.x[i]-eta.x[ind.low])
        /(eta.x[ind.low+1]-eta.x[ind.low])
    }
}
return(deri.eta.result)
}

valideta <- function(eta) TRUE
link <- paste("linkalpha(" , deparse(substitute(alpha)) , ",")", sep="")
structure(list(linkfun = linkfun, linkinv = linkinv,
mu.eta = mu.eta, valideta = valideta,
name = link), class = "link-glm")

# derivative of the inverse link function
inv.deri.grid.fn <- function(alpha){
  eta.grid.forderi <- seq(-10,10,by=0.01)
  inv.result.forderi <- inv.link.fn(alpha,eta.grid.forderi)
  diff.forderi<- diff(inv.result.forderi)
  deri <- diff.forderi[,2]/diff.forderi[,1]
  l.x <- length(deri)
  eta.x <- inv.result.forderi[(1:l.x),1]+diff.forderi[,1]/2
  ind.in <- which(deri!='NaN')
  deri.result <- cbind(eta.x[ind.in],deri[ind.in])
  return(deri.result)
}

# fit a glmm model for a given value of alpha
ctrl <- glmerControl(tolPwrss=1e-4)
inv.deri.grid.result <- inv.deri.grid.fn(alpha)
gm1 <- glmer(cbind(event,n-event) ~ 1+group+(1+group|study),
              family=binomial(link=linkalpha(alpha=0.5)), control=ctrl,
              data=data)
Bibliography


