Statistical Challenges in Individual Patient-Data Meta-Analyses of Binary Outcomes

Doneal Thomas

Master of Science

Department of Epidemiology, Biostatistics and Occupational Health

McGill University
Montreal, Quebec, Canada

June 2015

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Master of Science.

ACKNOWLEDGMENTS

I would first like to give thanks to God for granting me life, along with its many advantages. He has made it possible when things seems impossible, he has given me the knowledge of problem solving, the skills needed to pursue this degree.

Dr. Andrea Benedetti, my supervisor, for guiding me through my research, and also her help academically. I would like to express my deepest sincerity to her on the advice given. For that I am indebted to her.

My parents, Alban and Teresa Thomas, the greatest parents in the world. Words along cannot express my gratitude to them. I thank them for bringing me into this world, for their guidance through life, academically challenging me to achieve success and the instilment of life morals. In addition, I extend thanks to my siblings Dr. Daniel Thomas, Euthlyn Bernard and Daina Sylvan.

On a more personal note I would like to thank my wife Nerissa Phillip-Thomas and children D'Andre and D'Vonte Thomas, words cannot express my gratitude for everything they have done and sacrifice. I thank them for their understanding and support on this wonderful journey. I Love you very much!

I want to thank my friends that I have made during my academic years, life long friends I have, and the many people who have believed in me, your belief has played a driving force in my success. I wish you all God's continual blessings.

Finally, during my studies I received Graduate excellence award, merit fellowship and internal studentship. I am thankful for these funding for they have supported me through these years.

ABSTRACT

Meta-analyses (MA) based on individual patient data (IPD) are regarded as the gold standard and are becoming increasingly common, having several advantages over MA of summary statistics. These analyses are being undertaken in an increasing diversity of settings, often having a binary outcome. Parameter estimation of generalized linear mixed models (GLMMs), which are frequently used to perform inference on binary outcomes, is convoluted by intractable integrals in the marginal likelihood. Penalized quasi-likelihood (PQL) and adaptive Gauss-Hermite quadrature (AGHQ) are estimation methods commonly used in practice. However, few comparisons for the assessment of the performances of these estimation methods have been reported in the context of IPD meta-analyses (IPD-MA) with binary outcomes.

I considered as a first step to the thesis, a systematic review of the literature. In a previous systematic review of articles published between 1999-2001, the statistical approach was seldom reported in sufficient detail, and the outcome was binary in 32% of the studies considered. Here, we explore statistical methods used for IPD-MA of binary outcomes only, a decade later. 19 of the 26 MA used a one-step approach verses a two-step approach and random-effect logistic regression was the most common method for these binary outcomes, allowing the treatment effect to vary across studies. However, the estimation technique used in these studies (e.g. a GLMM estimated via PQL or AGHQ) was rarely reported.

Afterwards, via simulation studies, whose design is realistic for conducting IPD-MA of binary outcomes, to compare techniques to estimate multilevel models, and to address the concern of including trial-membership as fixed or random? The parameters of the one-step models were estimated using PQL and AGHQ while that of the two-step model were estimated via restricted maximum likelihood (REML) at the second step. Size and number of study, total sample sizes, variances and correlation in the random effects distribution were varied. The comparison is done in terms of bias, root mean square error (RMSE), numerical convergence, and coverage of interval estimates. The two-step and one-step (via PQL, and AGHQ) methods produced approximately unbiased pooled treatment effect estimates, although the manner in which PQL achieves this is an advantage. The AGHQ methods for estimating the random treatment effect variance performed better with respect to bias and coverage, but RMSE performed relatively poor in comparison with PQL for all data sizes and model misspecification.

ABRÉGÉ

Les méta-analyses (MA) de données individuelles de patient (IPD) sont considérées comme une approche de référence et deviennent de plus en plus communes puisqu'elles ont plusieurs avantages comparativement aux méta-analyses de statistiques sommaires. Le champ d'application de ces analyses est diversifié et relève souvent une réponse binaire. L'estimation des paramètres dans de modèles linéaires généralisés mixtes (GLMMs), qui sont souvent utilisés pour inférer sur des réponses binaires, est compliquée par l'évaluation d'intégrales insolubles dans la vraisemblance marginale. Les méthodes d'estimation quasi-vraisemblance pénalisée (PQL) et quadrature Gauss-Hermite adaptive (AGHQ) sont couramment utilisées dans la pratique. Cependant, peu de comparaisons sur l'évaluation de la performance de ces méthodes d'estimation ont été rapportées dans le contexte de méta-analyses de données individuelles de patient (IPD-MA) avec des réponses binaires.

La première étape de la thèse est une revue systématique de la littérature. Dans une précédente revue systématique d'articles publiés entre 1999 et 2001, la méthode statistique était rarement rapportée avec suffisamment de détails et 32% des études considérées reportaient une réponse binaire. Une décennie plus tard, nous explorons les méthodes statistiques utilisées seulement pour les IPD-MA avec des réponses binaires. 19 des 26 MA identifiées utilisaient une approche en une étape au lieu d'une approche en deux étapes et la regression logistique avec effets aléatoires était la méthode la plus commune pour ces données binaires, permettant à l'effet du traitement de varier d'une étude à l'autre. Cependant, la technique d'estimation utilisée dans ces études (e.g. un GLMM estimé via PQL ou AGHQ) était rarement rapportée.

Ensuite, des études de simulation dont la conception est réaliste pour appliquer une IPD-MA avec des réponses binaires sont réalisées pour comparer des techniques d'estimation de modèles à plusieurs niveaux et pour considérer l'inclusion de l'adhésion provisoire en tant qu'effet fixe ou aléatoire. Les paramètres de l'approche en une étape sont estimés avec PQL et AGHQ tandis que ceux de l'approche en deux étapes sont estimés via la vraisemblance maximale restreinte (REML) à la deuxième étape. La taille et le nombre d'études, les tailles d'échantillon totales, les variances et les corrélation de la distribution des effets aléatoires sont variés. La comparaison concerne le biais, l'erreur quadratique moyenne (RSME), la convergence numérique et la couverture des intervalles des estimés. Les approches en une et deux étapes (via PQL et AGHQ) produisaient des estimations combinées de l'effet du traitement approximativement non-biaisées, bien que la manière dont la méthode PQL produisait cette estimation soit avantageuse. La méthode AGHQ pour estimer la variance des effets aléatoires du traitement performaient mieux concernant le biais et la couverture, mais RSME performait relativement mal comparativement à PQL pour toutes les tailles de données et les mauvaises spécifications de modèle.

Preface

Format of thesis

This thesis includes two manuscripts presented in Chapters 4 and 5 I opted to write this thesis in the manuscript-based format that follows the requirement of McGill University regulations. The thesis is comprised of an introduction, a literature review, the two manuscript chapters and a final conclusion chapter. The manuscripts are connected through the preamble included for each.

Additional results not presented in the manuscripts, are presented in the appendix at the end of each chapter.

Contributions of authors

I led this work in writing, programmed the simulation study, preformed all the analyses, interpreted and discussed the results of both manuscripts and was a reviewer of the original manuscripts included in manuscript I. Both manuscripts have been co-authored by my thesis supervisor, Dr. Andrea Benedetti who contributed **in** developing the scope and statistical questions to address. My supervisor provided guidance, reviewed and corrected the thesis.

TABLE OF CONTENTS

LIST OF TABLES BY CHAPTER

CHAPTER 4

CHAPTER 5

LIST OF FIGURES BY CHAPTER

CHAPTER 4

CHAPTER 5

Chapter 1 Introduction

1.1 Rationale

Individual patient data meta-analyses (IPD-MA) are the gold standard of meta-analysis, having many advantages over conventional meta-analyses, particularly when the outcome is binary and modelled using the one-step method. Such models offer more possibilities of complex modelling, but empirically should perform comparably to the two-step methods in some situations.

Further, the methods used for the analysis of IPD-MA with binary outcomes show wide variability in practice. However, several statistical challenges remain to be investigated.

Chapter 2 Literature Review

2.1 Overview of Meta-analysis Techniques and the Role of IPD Analysis

DerSimonian and Laird define meta-analysis as “the statistical analysis of a collection of analytic results for the purpose of integrating their findings” [1]. Meta-analyses (MA) can be considered a formal method for pooling information from a wide variety of sources and sometimes can be used to develop a consensus within the research community [2]. The method is particularly useful in fields where both the number of studies and the need for synthesis of the information is important [1 ,3]. The role of MA in summarizing scientific literature has expanded as the number of published studies has increased [1].

The role of MA in summarizing randomized clinical trials (RCTs) has been comprehensively studied and broadly utilized in clinical practice [4]. However, the use of MA techniques for summarizing results from observational studies is a more recent phenomenon and somewhat poses several unique challenges. Observational studies are more likely to suffer from uncontrolled confounding than randomized clinical trials and often times these biases are impossible to rule out [5]. The issue is further compounded by the diversity in study designs, data collection methods, analytic techniques and non-standardized reporting of observational studies. Although MA restricted to RCTs are usually preferred to MA of observational studies, there is still a need for synthesizing evidence in areas that are not amenable to RCTs, to enable optimal decision-making [6].

Conventional MA usually pool aggregate summary statistics of studies (e.g. odds ratios, risk differences, rate ratios, means, proportions, etc.), extracted from journal articles or via contacting the study investigator. Aggregate data MA suffer from several limitations [7-10]. Differences in study design can make it difficult to justify pooling results and to actually carry out MA [8]. Studies may also use different research methods and different modeling procedures in their published results, creating difficulties in combining their results [6 ,10]. The number of available studies for a meta-analysis can diminish rapidly as these inherent differences are discovered, hence reducing the power of the analyses. Stewart and Tierney[9] also pointed out that if only summary statistics are presented in the literature, it can be impossible to perform certain types of analyses such as time-to-event and to pool effects that have been adjusted for different variables [6]. With aggregate data MA, it is difficult to estimate the effects of patient-level covariates on the treatment effect [11]. In the context of an aggregate-data MA, this is known as meta-regression and may use study level covariates or aggregated patient level information. Meta-regressions on patient-level characteristics are prone to aggregation or ecological bias, and to confounding from variables not included in the model [3 ,10 ,12]. Such an analysis must use average patient characteristics, and the bias occurs from the mistaken assumption that a statistical association between average patient variables across trials is equal to the association between the corresponding variables at the individual level [12].

2.2 Advantages of Individual Patient Data-Meta-Analyses over Aggregate Data Meta-Analyses

In order to overcome some of these problems, collaborative groups are increasingly collecting raw data for each patient in each study and performing what is known as individual patient data meta-analyses (IPD-MA) [13]. Stewart and Tierney [9] defined IPD-MA as “the central collection, validation, and reanalysis of ‘raw’ data from all clinical trials (or observational studies) worldwide that have addressed a common research question.” IPD-MA are considered to be the least biased method as compared to aggregated data MA and is termed the “gold standard” for addressing many problems associated with using data from published articles, and a few of the problems associated with synthesizing summary data [5 ,7].

Some of these advantages include having access to a complete and up-to-date dataset from each of the included studies on which to base analyses, being able to perform standardized statistical analyses across studies and being able to have consistent inclusion and exclusion criteria across studies [2]. IPD-MA permit the possibility of detailed statistical analyses including subgroup analyses and the ability to adjust for confounders within and between studies [2].

Furthermore, the benefits of IPD-MA over aggregate-data MA, according to Stewart and Tierney, are that a meta-analyst can reduce or eliminate publication bias by incorporating the results from unpublished studies [9]. That is, when some studies are more likely to be published than others, the literature that is available may provide misleading information. In addition, subgroup analyses can be performed on IPD data, different scales of measurements can be combined and alternative but related questions can be investigated [9].

Stewart and Parmar compared the two methods using data from the Advanced Ovarian Cancer Trials Group that investigated non-platinum drugs and platinum based chemotherapy for cancer treatment [5]. The analysis of the aggregated data (AD) MA gave a result of greater statistical significance (p-value=0.027 vs p-value=0.30) and an estimate of absolute treatment effect three times as large as the IPD-MA (7.5% vs 2.5%) [5]. The method of analysis contributed to this observed difference [5].

However, the results of AD-MA can coincide with that of IPD-MA [2]. If complete aggregate data extracted from published studies can be obtained, then a two-step IPD-MA (as discussed below) and an AD meta-analysis will yield similar results, conditional on the other factors (number of patients etc.) [14 ,15].

As discussed, IPD-MA offer several advantages over conventional aggregated data MA, however, strikingly, a standardized data analytic approach for IPD-MA does not exist, and has been noted as one of the drawbacks for IPD-MA [16]. In this work the focus is on IPD-MA of binary outcomes that arise frequently in practice.

2.3 Methods for Combining Effects: One- verses Two-Step

Two broad analytic approaches exist for combining the results from multiple studies in IPD-MA. One approach requires that the analyst conduct individual analyses for each study then generate summary statistics (such as a difference in means or log odds ratio and a standard error of the estimate) that would be published in the literature, and use classical MA approaches, such as, inverse-variance weighting, to synthesize these summary statistics. This is known as the two-step approach. A two-step analysis of IPD improves upon conventional aggregate-data MA due to the standardization of inclusion/exclusion criteria, exposure and outcome definitions and statistical analysis. However, it does not take full advantage of the richness of an IPD [17]. A one-step approach as described next, should confer even more advantages.

One-step methods use a single statistical model, while accounting for the clustering among patients in the same study, to estimate an overall effect. A one-step model also takes advantage of the ability to standardize elements of the analysis across studies [2 ,3 ,15]. However, a one-step model allows investigation of patient- and trial-level covariates [11]. In particular, this approach allows investigation of dose-response curves (e.g. allowing non-linearity), improves power for interactions and subgroup analyses [8 ,10 ,18], and allows control of confounding by patient- and study-level covariates in a much better way than a two-step approach [2 ,3 ,15]. A one-step approach also offers more flexibility to explore the differences that may exist between patients in the same study as well as across studies [19]. The problems of zero cells in some studies [12], usually excluding smaller studies are also addressed by using a one-step model. This approach allows studies with zero cells to provide some information [12].

Recently, some have suggested that the two-step and one-step approaches to analysis of IPD-MA produce similar results for meta-analyses of large randomized controlled trials [20]. However, another study showed that occasionally the one-step and two-step approaches lead to different conclusions about which factors are associated with the outcome [17]. The literature suggests that the one-step method is particularly preferable when few studies or few events are available as it uses a more exact statistical approach than reducing the evidence of MA to assume approximate normality [17]. Further, there is little agreement between the one- and two-step methods when interest lies in identifying patient-level characteristics that are related to treatment-effect (effect modification) [8 ,12]. One-step method will always improve on the power to adequately identify clinically moderate interactions over the two-step method [8].

2.4 Fixed verses Random Effects

Within IPD-MA, the study and/or the treatment may be specified as either a fixed- or random-effect; where the choice of method depends on the question to be answered.

Under the fixed-effects model we assume that the true effect size for all studies is identical, and the only reason the effect size varies between studies is by the play of chance, that is, within-study sampling error. The fixed-effects model may be used if (i) it is believed that all the studies included in the analysis are functionally identical and (ii) the goal is to compute the common exposure effect for the identified population, and not to generalize to other populations [21]. Typically, studies are not functionally identical -- the subjects or interventions differ in ways that may have impacted the results, and therefore one should not assume a common exposure effect. In these cases the random study- and/or treatment-effect models are more easily justified that the fixed effects model. The goal of the random effects model is not to estimate one true effect, but to estimate the mean of a distribution of effects. Conventionally the normal distribution has been used to accommodate the variation [21]. Additionally, the uses of the random effects models can be generalized to a range of scenarios.

As demonstrated by Straum [22] the mixed effect regression model structure is a very general framework that can incorporate random effects for the studies, treatments and covariates

[22 ,23], use data from both single and two-arm studies [24], and offer the ability to combine surrogate endpoints [25].

2.5 Heterogeneity in Meta-analysis

In any meta-analysis the point estimates of the exposure effect from the different studies being considered will almost always differ, to some degree. When exposure effects differ, but only due to sampling error (i.e. the true effect is the same in each study), the effect estimates are considered to be homogeneous; in other words, differences between estimates are random variations, and not due to systematic differences between studies. However, often the variability in the effect size estimates exceeds that expected from sampling error alone, that is, there is not just the same true underlying effect for each study, but "real" differences exist between studies. When it is present, the effect size estimates are considered to be heterogeneous and potential sources of heterogeneity should be explored.

Heterogeneity may be related to differences between included patient populations, details of the interventions and co-interventions, and measurement and definition of the outcome [26]. Large heterogeneity demands some action, e.g. stratifying by trial features or meta-regression to identify sources of heterogeneity [27]; and suggests using a random-effects analysis, though some argue it precludes pooling effects [27]. Various methods exist for assessing and quantifying these differences between the study estimates. These range from simple graphical assessments (e.g. the forest plot [28]) to complicated formal statistical tests and estimation methods, such as I^2 [27 ,29]. While the I^2 is the standard, some have argued that between-study variability may be best described simply by τ^2 [29 ,30].

2.6 Overview of generalized linear mixed models

The one-step random effects model for binary outcome data described above is a form of generalized linear mixed model (GLMM). GLMMs extend generalized linear models (GLMs) by adding normally distributed random effects to the linear predictors of a GLM, to account for correlation among the responses [31]. The GLMMs are special cases of random effects models, as in some cases the random effects are not normally distributed [32]. Random effects can be interpreted as reflecting the natural heterogeneity across studies due to unmeasured factors or heterogeneity not captured by covariates included in the linear predictor [33]. For example, in patients with a suspected presence of Deep Vein Thrombosis (DVP)- IPD of patients can be collected and various studies can aim at estimating which candidate factors are associated with the outcome. These factors would induce correlation in responses for a patient within study, as well as, heterogeneity in responses of patients between studies (differences in geographic local, setting and time) [17].

To analyze IPD-MA with a dichotomous outcome measure under the fixed effects GLMs, the following notation will be adopted: y_{ij} is the observed outcome for an individual, coming from a binomial distribution with parameter $P\left(Y_{ij} = 1\right) = \pi_{ij}$ and a denominator of 1. If π_{ij} is the true response probability for the jth individual in the ith study where $i = 1, \dots, K$, then

$$logit\left(\pi_{ij}\ \right) = \beta_0 + \beta_1 z_{ij}$$

where $\boldsymbol{\beta_0}$ is the set of fixed study effects which represent the average log odds among the untreated subjects in each study, $\boldsymbol{\beta_1}$ is the pooled exposure effect (log odds ratio) of an intervention as compared to control and z_{ij} is an indicator variable for treatment assignment.

Under the same methodological framework, the general form of the random effects model with random study effects includes the effects b_{0i} of study on the log –odds as well as the effects b_{1i} of study on treatment effects:

$$logit(\pi_{ij}\ ) = \beta_0 + b_{0i} + (\beta_1 + b_{1i})z_{ij}$$

$$\text{where } \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}$$

When modeling treatment- and study-effects as random, we are implicitly modeling the variance-covariance matrix associated with parameter [34]. If $cov(b_{0i}, b_{1i})$ is assumed to be zero, the between-study variance of the log-odds across control groups is modeled by σ^2, while that across treatment groups is modeled by $\sigma^2 + \tau^2$ [34]. The variation across study for control groups is thereby forced to be less than or equal to the variation across study for treatment groups; this assumption may well be inappropriate [34]. Therefore, the covariance between the random intercepts and slopes should be modeled.

A GLMM can be specified as follows. As before, let y_{ij} represent the i^{th} individual in the j^{th} study with $i = 1, \dots, n_j$ and $j = 1, \dots, K$. It is assumed, conditional on random effects, b_j, the responses Y_{ij} are independent with mean:

$$logit\ E(y_{ij}|b_j) = x_{ij}^T\beta + \ z_{ij}^T b_J,$$

where $logit(x) = \log\left(\frac{x}{1-x}\right)$ and $\boldsymbol{b_j}$ have a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{G} = \boldsymbol{G(\gamma)}$ [31]**.** If all elements of the covariance matrix $\boldsymbol{G}$ equal 0, all of the random effects equal 0 and the GLMM is reduced to a GLM. The appropriateness of the normality assumption for the random effects is difficult to verify [31].

2.7 The problem of the likelihood inference for GLMMs

The maximum likelihood estimates of the GLMM parameters are obtained by integrating out the unobserved random effects $\boldsymbol{b_i}$ from the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{b_i}$**.** The marginal likelihood is defined as

$$L(\beta, G, \phi) = \prod_{i=1}^{K} \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|b_i, \beta, \phi\right) f(b_i|G)\, db_i\,,$$

where $f_{ij}\left(y_{ij}\middle|b_i, \beta, \phi\right)$ is the conditional density of the outcome Y_{ij} and $f(b_i|G)$ the density of the random effects, that is, a multivariate normal distribution.

The problem with the expression for the likelihood is that it is generally not analytically tractable, and so the likelihood cannot be expressed in closed form, except in special cases such as linear mixed models (LMMs). A variety of approaches have been proposed to circumvent this difficulty, including approximate likelihood approaches based on the use of the Laplace approximation, such as penalized quasi-likelihood (PQL) introduced by Breslow and Clayton [35], numerical approaches, such as Gauss-Hermite quadrature (GHQ) [36], and approaches based on the use of Monte Carlo methods, such as modern Bayesian approaches implementing

Markov Chain Monte Carlo (MCMC) techniques [37]. These methods will be explored in detail in the following section.

2.8 Estimation methods for GLMMs

The main difficulty with GLMM estimation is the presence of high dimensional integrals with no analytics solutions. A detailed overview of Adaptive Gauss-Hermite quadrature (AGHQ), PQL and some of the relative merits and disadvantages of each approach raised in the literature are provided below.

2.8.1 Penalized Quasi-likelihood

PQL is the most well known Laplace approximate likelihood approach for GLMMs. It is an iterative algorithm for solving a GLMM, which is similar to the iterative least squares method for solving a GLM [35 ,38].

The simplest derivation for the PQL technique is based on a first order Taylor series approximation [39 ,40]. If the GLMM is decomposed into

$$y_{ij} \approx \mu_{ij} + \epsilon_{ij} = h(\eta_{ij}) + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N\left(0, \phi v(\mu_{ij})\right),$$

where $h = g^{-1}$ is the inverse link function, then an expansion around the current iteration value of the linear predictor, $\boldsymbol{\eta}_{\boldsymbol{ij}}^{(\boldsymbol{k-1})} = \boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{T}}\boldsymbol{\beta}^{(\boldsymbol{k-1})} + \boldsymbol{z}_{\boldsymbol{ij}}^{\boldsymbol{T}}\boldsymbol{b}^{(\boldsymbol{k-1})}$**,** gives the pseudo-data defined as

$$y_{ij} \approx h\left(\eta_{ij}^{(k-1)}\right) + h'\left(\eta_{ij}^{(k-1)}\right)\left(\boldsymbol{x}_{\boldsymbol{ij}}^{\boldsymbol{T}}\boldsymbol{\beta} + \boldsymbol{z}_{\boldsymbol{ij}}^{\boldsymbol{T}}\boldsymbol{b} - \eta_{ij}^{(k-1)}\right) + \epsilon_{ij}.$$

The PQL method is based on two analytic approximations: a Taylor series approximation and a probabilistic approximation of the normality assumption [39]. The expression given above follows the form of a LMM for the pseudo-data. The maximization of the likelihood for LMMs does not suffer from the problem of intractable integrals. The iterative steps of the PQL method alternates between fitting a LMM to the pseudo-data to obtain new estimates for $\boldsymbol{\beta}, G, \phi$ and $\boldsymbol{b}$, then updates the pseudo-data with the new estimates [35].

Furthermore, the PQL method is considered to be a computationally efficient way of fitting a wide variety of GLMMs as compared to numerical methods, such as AGHQ. PQL can be utilized to estimate parameters of GLMMs with several random effects, with more than two levels. The biggest known problem with the PQL approach, highlighted in previous literatures, is the potential for large estimation biases for some GLMMs, such as for binary data with small cluster sizes [38 ,41]. The bias in the PQL estimates has led to the development of a series of modifications and proposals: correction of PQL [41 ,42], modified Laplace approximation [43 ,44], and higher order Laplace approximations [45]. The list is long, and one can find thorough reviews of these developments in McCullagh and Searle [31], Demidenko [46], Hedeker and Gibbons [47] and Lee et al. [48].

2.8.2 Adaptive Gauss-Hermite Quadrature

AGHQ is the current favored competitor to PQL, which approximate the likelihood by numerical integration via a weighted sum of the GHQ approximation [36]. Numerical integration proceeds by GHQ formula

$$\int_{-\infty}^{\infty} h(v)e^{-v^2}dv \cong \sum_{q=1}^{m} h(x_q)w_q,$$

where h is a smooth function, $\boldsymbol{x_q}$ are the quadrature points that are the roots of the mth order Hermite polynomial with corresponding weights $\boldsymbol{w_q}$**,** both of which are available from standard references [49]. The larger the m, the number of quadrature points, the better the approximation. When one quadrature point is used, AGHQ is equivalent to the Laplace approximation [36]. The integrand of the contribution of study *i* to the marginal likelihood can be transformed to the form of GHQ formula above by substituting the linear transformation $v_i = (2G)^{-1/2}b_i$ with $dv_i = (2G)^{-1/2}db_i$, considering a random intercept b_i only model and G the variance of the random intercept distribution

$$f_i(\boldsymbol{y_i}|\boldsymbol{\beta}, G, \phi) = \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i, \boldsymbol{\beta}, \phi) f(b_i|G)\, db_i$$

$$= \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i, \boldsymbol{\beta}, \phi) \frac{e^{-b_i/2G}}{(2\pi G)^{1/2}} db_i$$

$$= \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i, \boldsymbol{\beta}, \phi) \frac{e^{-((2G)^{-1/2}b_i)^2}}{\sqrt{\pi}} db_i$$

$$= \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i = (2G)^{1/2}v_i, \boldsymbol{\beta}, \phi) \frac{e^{-v_i^2}}{\sqrt{\pi}} dv_i$$

$$= \int_{-\infty}^{\infty} h^*(v_i)e^{-{v_i}^2} dv_i,$$

where $h^*(\cdot)$ is the conditional distribution for the vector of responses given the random effects b_i, divided by a normalizing constant $\sqrt{\pi}$. The estimator obtained by maximizing the likelihood approximated in this way is called the non-adaptive Gauss-Hermite quadrature (NGQ) estimator.

Naylor & Smith [50] and Liu & Pierce [51] suggested an improvement to standard GHQ. While GHQ approximates the likelihood by choosing optimal subsets at which to evaluate the integrand, AGHQ uses further information gained from an initial parameterization to increase precision [36]. However, there is a trade-off with respect to precision and speed. AGHQ gets exponentially slower as the dimension of the random effects increases, to an extent where the procedure is not feasible for 2 or 3 random effects [36].

For simplicity, the AGHQ method is described here for a univariate integral $\int_{-\infty}^{\infty} f(x)dx$. Let $\hat{\mu}$ be the mode of $f(x)$ or the mean of a variable with probability density function (pdf) proportional to $f(x)$, and let $\hat{\sigma}^2$ represent either the estimated curvature of $f(x)$ at the mode $\hat{\mu}$ or the variance of a variable with pdf proportional to $f(x)$. Consider an integrand including the product of a function and an arbitrary normal density

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{f(x)}{\varphi(x;\hat{\mu},\hat{\sigma}^2)}\varphi(x;\hat{\mu},\hat{\sigma}^2)dx = \int_{-\infty}^{\infty} h(x)\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}e^{-(x-\hat{\mu})/2\hat{\sigma}^2}dx,$$

where $h(x) = f(x)/\varphi(x;\hat{\mu},\hat{\sigma}^2)$ and $\varphi(x;\mu,\sigma^2)$ is a normal pdf with parameters μ and σ^2. The application of GHQ to this integral implies the re-parameterization of the function $h(\cdot)$ at quadrature point $x_q = \hat{\mu} + \sqrt{2}\hat{\sigma}t_q$, where t_q are the roots of the mth order Hermite polynomial. Applying the change of variable $t = (x - \hat{\mu})/\sqrt{2}\hat{\sigma}$, the expression becomes

$$\int_{-\infty}^{\infty} h(\hat{\mu} + \sqrt{2}\hat{\sigma}t)\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}e^{-t^2}\sqrt{2}\hat{\sigma}dt$$

$$= \int_{-\infty}^{\infty} h(\hat{\mu} + \sqrt{2}\hat{\sigma}t) \frac{1}{\sqrt{\pi}} e^{-t^2} dt$$

$$\approx \sum_{q=1}^{m} \frac{w_q}{\sqrt{\pi}} h(x_q)$$

In short, the adaptive process takes into account the properties of the entire integrand. The method scale and translate the quadrature locations to place them under the peak of the integrand. In this way, the position of the quadrature points may vary from study to study.

2.9 Review of previous simulation studies on the Estimation Methods

The literature on likelihood-based methods for the maximum likelihood estimation of the parameters for GLMMs is vast. Diaz [52] compared the PQL approach with a special case of the Laplace approximation for the logit-normal model in a cluster randomized trial setting. He found that the numerical Laplace method might have reduced the bias, but at the expense of huge variances.

Callens and Croux [53] compared the performance of PQL, to non-adaptive and adaptive Gaussian methods for correlated binary outcomes. In this paper it was concluded that (i) PQL is a much faster estimation method, (ii) AGHQ is preferred than non-adaptive and (iii) although PQL suffers from larger bias in the parameter estimates (confirming to findings in the literature), it performs better in terms of mean square error.

Few comparisons between the various estimation methods have been reported in the context of IPD-MA with binary outcomes – that is few clusters, imbalanced cluster sizes, large inter-study heterogeneity, small exposure effects and an interest in the variance parameter of the random treatment effect [17 ,34]. In addition, several simulation studies limit their attention to simple models with only random intercepts. The performances of the random effects models that include both a random intercept and a random slope, or only a random slope are far less reported [34]. The simulation studies presented in manuscript II assess the performance of likelihood-based methods (PQL and AGHQ) for fitting stratified- and random-intercepts with an interest in the inter-study heterogeneity of the treatment effect in realistic settings common when IPD-MA are performed with binary outcomes.

Furthermore, there is little discussions of methodology for IPD-MA with binary outcomes in the literature, several concentrate either on general practicalities of IPD reviews or advance multilevel modeling techniques [16]. The lack of a standardized data analysis plan and guidance has led to various differences in the approach and conduct. To identify areas in need of guidance and further research, the systematic review presented in manuscript I investigate the statistical approach taken to analyze recent IPD-MA with binary endpoints.

Chapter 3 Objectives

Two studies were undertaken whose manuscripts form this dissertation. The general objective was to further the development of bio-statistical methodology for the analysis of individual patient data meta-analyses (IPD-MA) with binary outcomes. The specific objectives were as follows:

1) To systematically investigate statistical methods used to analyze IPD-MA with binary outcomes and update a previous review of these methods.

Despite the many advantages in carrying out IPD-MA over aggregate data meta-analyses (i.e. standardizing the statistical analysis across studies, assessing patient-level data, examining interaction terms etc.), the obtainment of IPD as well as the cost associated can be prohibitive. Further, the wide range of methods used for the analysis of IPD-MA (one- vs. two-step, fixed- vs. random-effects etc.) and the lack of a standardized protocol for the data analysis is a limitation.

2a) To compare various performance aspects of two estimation procedures (Penalized quasi-likelihood and Adaptive Gauss-Hermite quadrature) for generalized linear mixed models (GLMMs) in the unique setting of IPD-MA under a one-step approach, to each other and the two-step method.

More specifically, the performance of penalized quasi-likelihood (PQL) and adaptive Gaussian Hermite quadrature (AGHQ) was investigated via simulation study. Performance assessment was via: bias, mean square error, coverage and numerical convergence, of the pooled treatment effect and the inter-study heterogeneity. The effects of number of studies, number of patients per study, variability of study size and heterogeneity of the treatment effects across studies was investigated.

2b) To investigate the assumption and implication of modelling the study-effects as fixed or random within GLMMs.

It is important to investigate different sources of heterogeneity in performing a meta-analysis. Some of this variability cannot only be explained by the heterogeneity of the treatment effects across studies but also by heterogeneity in the baseline odds. In fitting a fixed study-effect, a dummy variable is estimated for each study included, reducing the information available for estimating any one-model parameter.

Preamble to Manuscript I

The first part of this thesis is a systematic review of the methodology for individual patient data meta-analysis (IPD-MA). A previous paper reviewed methods used in practice for meta-analysis of IPD from randomized trials. That paper reviewed 44 articles published during 1999–2001, of which 14 considered a binary outcome and found that the two-step approach was used about two-thirds of the time [16]. They also found that few details were provided. Further, the lack of guidance has led to vast variation in the approaches.

Over the intervening years, generalized linear mixed models have become increasingly commonplace-they are used more frequently in the medical research literature, and are available in most statistical software packages. Given this, we were interested in updating the previous review.

The second aspect considered in this manuscript was to investigate the analytical approaches to IPD-MA with binary outcomes. When there is substantial variation across studies, various random-effects meta-analysis models are possible that employ a one-step or two-step method. Empirical comparisons are few between the methods; therefore, it was interesting to compare these approaches and report, which is frequently employed in practice.

This article was submitted and published in the BMC Medical Research Methodology journal.

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Systematic review of methods for individual patient data meta- analysis with binary outcomes

Doneal Thomas[1], Sanyath Radji[4] and Andrea Benedetti[1,2,3,5*]

Abstract

Background: Meta-analyses (MA) based on individual patient data (IPD) are regarded as the gold standard for meta-analyses and are becoming increasingly common, having several advantages over meta-analyses of summary statistics. These analyses are being undertaken in an increasing diversity of settings, often having a binary outcome. In a previous systematic review of articles published between 1999–2001, the statistical approach was seldom reported in sufficient detail, and the outcome was binary in 32% of the studies considered. Here, we explore statistical methods used for IPD-MA of binary outcomes only, a decade later.

Methods: We selected 56 articles, published in 2011 that presented results from an individual patient data meta-analysis. Of these, 26 considered a binary outcome. Here, we review 26 IPD-MA published during 2011 to consider: the goal of the study and reason for conducting an IPD-MA, whether they obtained all the data they sought, the approach used in their analysis, for instance, a two-stage or a one stage model, and the assumption of fixed or random effects. We also investigated how heterogeneity across studies was described and how studies investigated the effects of covariates.

Results: 19 of the 26 IPD-MA used a one-stage approach. 9 IPD-MA used a one-stage random treatment-effect logistic regression model, allowing the treatment effect to vary across studies. Twelve IPD-MA presented some form of statistic to measure heterogeneity across studies, though these were usually calculated using two-stage approach. Subgroup analyses were undertaken in all IPD-MA that aimed to estimate a treatment effect or safety of a treatment,. Sixteen meta-analyses obtained 90% or more of the patients sought.

Conclusion: Evidence from this systematic review shows that the use of binary outcomes in assessing the effects of health care problems has increased, with random effects logistic regression the most common method of analysis. Methods are still often not reported in enough detail. Results also show that heterogeneity of treatment effects is discussed in most applications.

Keywords: Individual patient data, Meta-analysis, Random effects, Systematic review, Heterogeneity, One-stage

Background

A meta-analysis (MA) attempts to synthesize the results from various distinct studies. The goal is to summarize the evidence for a particular statistical measure of interest, such as a risk difference or odds ratio. It is an especially important tool in clinical practice and medical research, where evidence-based information is preferred [1].

Individual patient data (IPD) MA are the gold standard of meta-analysis. In an IPD-MA line-by-line patient data are collected from the relevant studies, rather than just the measure of effect as in a standard aggregate data (AD) MA. This permits researchers to define exposures and outcomes consistently across studies, and to analyze them more similarly (e.g. adjusting for the same confounders), which may minimize heterogeneity [2,3].

For IPD-MA, two broad analytic strategies (one- and two-step approaches) are possible; both preserve the clustering of subjects within studies, comparability of study arms, and both may be either fixed or random. A fixed effects analysis assumes that the estimated effect is the same across all studies, while a random effects analysis assumes

* Correspondence: andrea.benedetti@mcgill.ca
[1]Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montreal, Canada
[2]Department of Medicine, McGill University, Montreal, Canada
Full list of author information is available at the end of the article

that the estimated effect varies across studies due to differences in patient populations, study procedures, etc [1,4].

A two-step approach first analyzes each study separately and as identically as possible, and then uses standard meta-analytic techniques to pool the measure of interest. The well-known random effects method of Der Simonian and Laird is frequently used in the second step of a two-step IPD-MA approach [1].

One step approaches use one statistical model while accounting for the clustering among patients in the same study, to estimate an overall effect. A one step model also takes advantage of the ability to standardize elements of the analysis across studies, but offers more flexibility to explore the differences that may exist between patients in the same study as well as across studies [2,3,5]. In particular, a one-step approach allows better control of confounding by patient- and study-level covariates, improves power for detecting interactions and subgroup analyses, as well as avoids and reduces the potential for ecological bias that may occur if group level information is included in the analysis [6,7].

In conventional AD-MA, it is difficult to estimate the effects of patient-level covariates on the treatment effect [8,9]. In the context of an AD-MA, this is known as meta-regression and may use study level covariates or aggregated patient level information. Meta-regressions are prone to ecological bias, and to confounding from variables not included in the model [5,6,9] and may have limited power. IPD-MA have higher power than meta-regression to detect the effect of an interaction between covariates and treatment, and are preferable when the interest is in estimating interactions with patient-level covariates [9-11].

Importantly, IPD-MA are not prone to ecological bias if inferences about individuals are not based on aggregated data and model misspecification is evaded [6]. For these reasons, and others, IPD-MA are considered the gold standard of meta-analysis, despite the complexity and cost of collecting the data, and are published with increasing frequency [2].

Despite the many advantages, the wide range of methods used for analysis of IPD-MA and the lack of a standardized data analysis plan is a serious drawback [12,13]. A previous review of methods used in practice for IPD-MA, reviewed 44 articles published during 1999–2001, of which 14 considered a binary outcome [13]. That review found that the two-step approach was used about two-thirds of the time [13].

The aim of this systematic review is to update that report, nearly a decade later when random effects models have been well integrated into other areas of health research, are readily available in many software packages and computing power is also up to the challenge. Our objective was to investigate the statistical approach taken to analyze IPD-MA with binary outcomes. In particular, we were interested in (i) whether two-stage or one-stage approaches were more common; (ii) how heterogeneity was investigated and reported; and (iii) if a one step approach was used, were intercepts permitted to vary across primary studies considered as random.

Methods

Eligibility criteria for included studies were articles published in 2011 that reported results of an individual patient data meta-analysis for a binary outcome and were indexed in PUBMED or Medline. We believed that this would provide a good overview of the methods currently used for analysis of IPD-MA. We performed the search in June 2012.

We searched in PUBMED and MEDLINE for articles published between January 1, 2011 and December 30, 2011. The search terms used were “meta analysis” and (“individual patient data” or “ipd” or “patient level” or “individual participant” or “integrated analysis”). The titles and abstracts of these articles were reviewed to ensure that they reported results of an IPD-MA.

For the full text review, a standardized form was filled independently by two reviewers (SR, DT). Discordant entries were resolved by a third reviewer (AB). The data we collected from each article included: the reason for performing an IPD-MA, the goal of the IPD-MA, the types of studies collected, the number of studies sought and retrieved; the number of patients sought and retrieved; the type of outcome (e.g., binary, time-to-event or continuous); the method of analysis for the primary outcome and whether the analytic approach was one-stage or two-stage; whether intercept and/or the treatment effect were allowed to vary across studies (fixed or random effects); how heterogeneity was quantified, addressed and reported; the method of analysis of covariates: whether by one- or two-stage methods; methods for study- or patient-level covariates; and, whether subgroup analyses were performed (See Additional file 1: Table S1). For this review, we have considered only those articles which used a binary outcome.

We present descriptive analyses only.

Results

A total of 111 articles were returned from our search strategy. The titles and abstracts of these articles were reviewed to ensure that they reported results of an individual patient data meta-analysis. On this basis, 56 were selected for full text review. Articles excluded did not report results from an individual patient data meta-analysis (See Figure 1).

Twenty-seven articles presented time-to-event outcome data, 2 presented continuous outcome data and only one article had a count outcome. We focus on the 26 articles that presented results using a binary outcome.

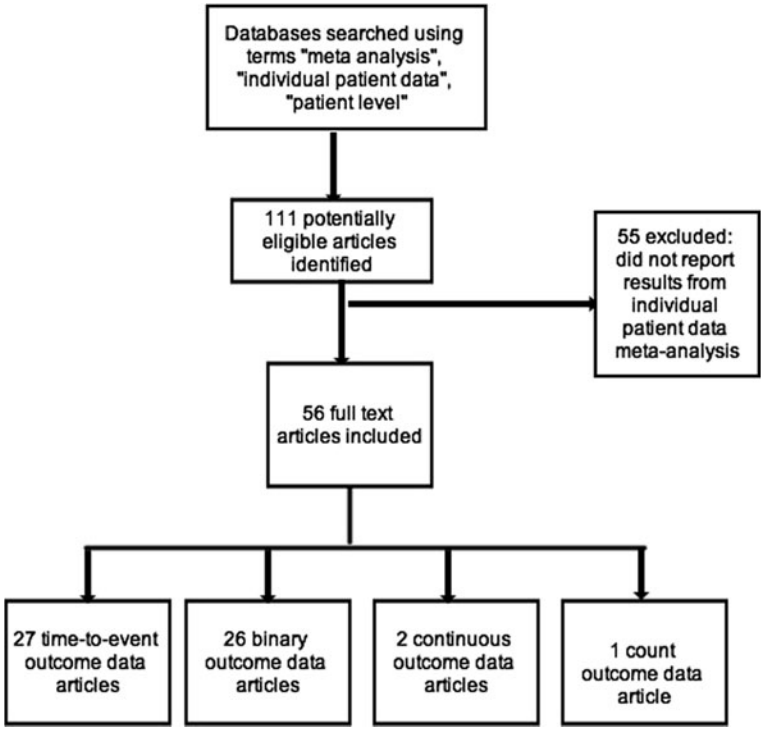
Databases searched using terms "meta analysis", "individual patient data", "patient level"
111 potentially eligible articles identified
55 excluded: did not report results from individual patient data meta-analysis
56 full text articles included
27 time-to-event outcome data articles
26 binary outcome data articles
2 continuous outcome data articles
1 count outcome data article

Figure 1 Flowchart of the inclusion of Individual patients data meta-analyses.

Among these 26 studies, the goals of the study were to estimate diagnostic accuracy (5, 19%) [14-18]; to estimate a treatment or exposure effect (14, 53%) [19-32], to identify predictors of an outcome (4, 15%) [23,33-35], to investigate safety of a treatment (3, 12%) [32,36,37], or other reason or goal not specified (2, 8%) [38,39]. (Note that percentages may not total to 100, because more than one goal was possible) (See Table 1).

Over half of IPD-MA (15/26) included only randomized control trials while the other IPD-MA included only observational studies. IPD-MA that included observational studies had a different profile in terms of goal with a greater proportion of studies that aimed to estimated diagnostic accuracy, and fewer IPD-MA that aimed to estimate the effect or safety of a treatment (See Table 1).

Why IPD?

When carrying out an IPD-MA, there are several advantages to be gained from this approach over aggregated data meta-analyses. The main reasons for adopting the IPD method reported for these 26 articles are summarized in Table 2. Half the studies included in our review cited subgroup analyses as the reason for conducting the IPD-MA.

Numbers of studies and patients

Figures 2 and 3 present the number of studies and number of patients included in the IPD-MA, respectively. More than 90% of the meta-analyses presented results for both the number of studies and patients obtained and sought. The median number of studies was 12, with interquartile range 6–18. The number of studies obtained in

Table 1 Goal of study, overall and stratified according to whether the IPD-MA included only randomized controlled trials, or included both randomized controlled trials and observational studies[1]

Reason	Included only randomized controlled trials (n = 15) N (%)	Included observational studies (n = 11) N (%)	Overall N (%)
To estimate a treatment effect	10 (67%)	3 (27%)	13 (50%)
To investigate safety of a treatment	2 (13%)	1 (9%)	3 (12%)
To estimate diagnostic accuracy	1 (7%)	4 (36%)	5 (19%)
To identify predictors	1 (7%)	3 (27%)	4 (15%)
Other/Unclear	2 (13%)	1 (9%)	3 (12%)

[1] Numbers may not total to 100% because some IPD-MA had more than one goal.

Table 2 Reasons provided to support conducting an IPD[1]

Reason	N (%)
To perform subgroup analyses	13 (50%)
To improve consistency across studies (in terms of inclusion criteria, outcome definition, etc.)	4 (15%)
To consider other outcomes	4 (15%)
To adjust for confounding variables	1 (4%)
To estimate diagnostic accuracy	5 (19%)
To identify predictors of an outcome	2 (8%)
Unclear	6 (23%)

[1]Percentages do not total to 100 because some studies reported more than one reason for conducting an IPD-MA.

the 26 meta-analyses ranged from about ten publications with fewer than ten studies, to five with more than twenty studies.

More variation was observed in the number of patients obtained, with median and inter-quartile range of 2964 and 679–4291 respectively (See Figure 3). Three meta-analyses had more than 10,000 patients and nine had fewer than 1000 patients.

Figure 4 shows the percentage of patients sought for which the full data were obtained. Sixteen (62%) meta-analyses obtained 90% or more of the total number of patients. Of these, eleven (69%) publications obtained information on all of the patients sought. The median of the 16 IPD-MA was 3430 with IQR of 908–6500 patients.

Statistical methods

Although many studies reported results for more than one outcome, here, we focus on the methods used to analyze the binary outcome. A majority of analyses concentrated on mortality or a dichotomized scale for the binary outcome. Most analyses used a one-stage method to pool the overall effect (69%) in the 26 IPD-MA for binary outcomes (Table 3). In those papers that used the one stage approach, usually all patient data from these studies were combined in a generalized linear mixed model (GLMM), accounting for the clustering among patients from the same study by including random study and or treatment effects. In general, few details were provided, and information often had to be inferred based on the results presented.

Among the 19 one-stage analyses, logistic regression was the most frequent technique employed. Ten of these IPD-MA used a random effects analysis. However, in 5 of these it was not clear whether intercepts, treatment effects or both were allowed to vary across studies. In the remaining 5 IPD-MA, 2 allowed both intercepts and treatment effects to vary, 1 allowed only the treatment effect to vary, and 2 allowed only the intercepts to vary. In general, little justification was offered for these choices. None specified the estimation method (e.g. penalized quasi-likelihood (PQL) [40] or adaptive Gaussian Hermite quadrature [41], etc.) used.

A fixed effects one-stage approach was used in 9 IPD-MA. Of these, 5 IPD-MA seemed to ignore clustering of subjects by study completely, and pooled all subjects together.

Two-stage methods were used in 6 of 26 studies reviewed. Of these, three studies used random effects for the treatment. One study initially used a Der Simonian Laird approach, but due to very low estimated heterogeneity, used a fixed treatment effect. The Cochrane-Mantel-Haenszel two-stage approach was used in one study, where no indication of heterogeneity across studies was found.

Heterogeneity

Most IPD-MA (n = 20) explicitly quantified heterogeneity across included studies. (See Table 4) The most frequently used measures were the Q statistic and I^2 [42], which were used in 12 studies. In five studies, other measures of heterogeneity were reported, such as the estimated variance from the random effects model or the inclusion of an interaction term in a model. It was unclear if any measure of heterogeneity was used in 6

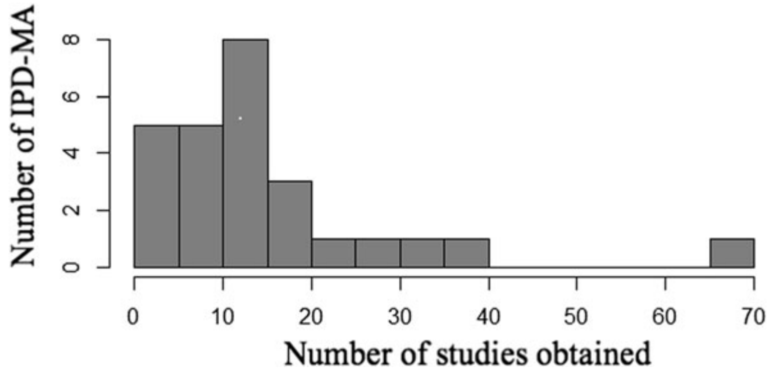

Figure 2 Number of studies from which IPD were obtained.

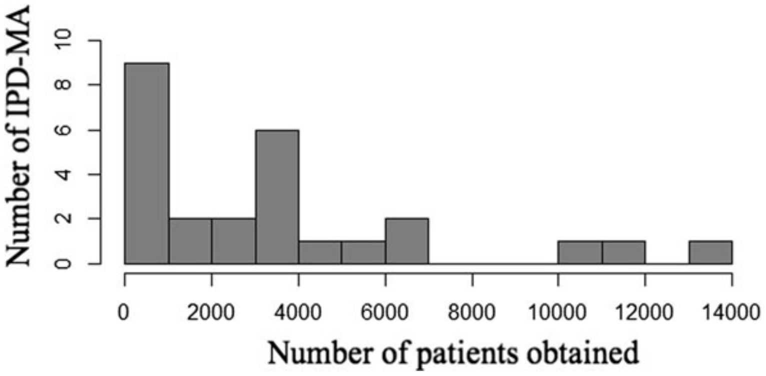

Figure 3 Number of patients from which IPD were obtained.

studies. In these studies no report or quantification of heterogeneity was presented. Two studies used multiple estimates to quantify heterogeneity; these estimates were the I^2 and Q statistics and the Breslow-Day and Q statistic [30]. Seven studies used a one step approach but reported measures of heterogeneity based on a two-step model, while the other studies used various techniques to assess and report heterogeneity.

Covariates

Covariates were used in three ways: (i) to assess subgroup effects; (ii) to adjust a treatment effect for possible confounders; and (iii) to identify predictors of an outcome.

Among the 16 studies where the goal of the IPD-MA was to estimate a treatment effect or the safety of a treatment, all considered subgroup analyses. Among studies that reported the number of subgroups considered, the median number of subgroups investigated was 2.5, with a range from 1–15. In all but one case, subgroups were formed by using categorical variables or categorizing a continuous variable. In one study, an interaction between the treatment and a continuous or ordinal risk score was evaluated. The subgroups investigated were based on patient-level characteristics in 13 IPD-MA, and on both patient- and study-level characteristics in 3 IPD-MA.

Among the studies that used a one-stage approach, 9/10 included interaction terms in the model, and presented stratum specific estimates as well as a p-value for the interaction. Among studies that used a two stage approach, 5/6 presented the stratum specific effect estimates, and 5/6 presented a p-value for the interaction. In two cases this p-value was calculated as described in [43].

Among the 3 IPD-MA that included observational studies and aimed to estimate a treatment effect or safety, all three adjusted for potential patient-level confounders. One of these studies used a two-step approach first adjusting for confounders in each study separately then pooling the adjusted effect estimates. Among the IPD-MA that only included randomized trials, and aimed to estimate a treatment effect or safety (n = 13), only 2 adjusted for patient level confounders. They did so by including them in a one stage model.

Finally, of the four IPD-MA that aimed to identify predictors of an outcome, three included observational studies.

Missing data

While there are a number of approaches that could be taken to deal with missing data, 16/26 IPD-MA did not report how missing data were handled. Three studies used

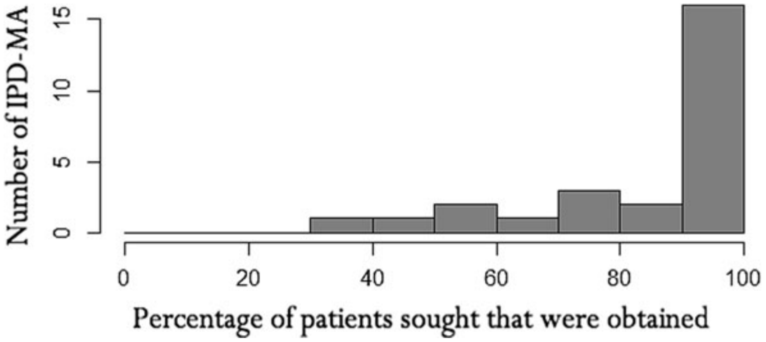

Figure 4 Percentage of patients sought that were obtained.

Table 3 Statistical analysis method categorized by overall strategy among 26 IPD meta-analyses of binary outcomes

Analytic approach[1]		n/N (%)
One stage approach (n = 19)		
Ignored clustering by study	Logistic regression	5/19 (26%)
Fixed effects	Logistic regression	4/19 (21%)
Random effects	Logistic regression	10/19 (52%)
	Fixed study effect with random treatment effect[2]	1/10 (10%)
	Random study effect with fixed treatment effect[2]	2/10 (20%)
	Random study effect with random treatment effect[2]	2/10 (20%)
	Unclear[1]	5/10 (50%)
Two stage approach (n = 6)		
Fixed effects	Unspecified	2/6 (33%)
	Cochrane-Mantel-Haenszel	1/6 (17%)
Random effects	Der Simonian Laird	2/6 (33%)
	Unspecified	1/6 (17%)

[1] It was unclear from one article [38] which approach was used.
[2] Among studies that used random effects logistic regression, where the intercepts and/or treatment effects allowed to vary across studies.

multiple imputation and two studies used single imputation. The remaining studies used a variety of other approaches to dealing with missing data including excluding subjects with missing data, or excluding variables with too much missing data, or it was unclear what approach was taken.

Discussion

In this paper, we reviewed a sample of published individual patient data meta-analyses where the primary outcome was dichotomous, focusing on the statistical approach taken and results reported. To identify relevant articles in our review, we used a thorough search strategy and assessed 26 IPD MA articles published in the year 2011 that presented results for a binary outcome. It is possible that some relevant papers that reported the results of IPD MA with binary outcomes and were published in 2011 have been missed or excluded unintentionally, but these would be unlikely to differ substantially methodologically than those included. Two reviewers extracted all information independently and a third reviewer resolved conflicts. It might also be possible due to the lack of sufficient details to distinguish the methods used, that methods were incorrectly classified since the precise method used was sometimes inferred.

Table 4 Statistic used to measure heterogeneity among studies in the 26 IPD meta-analyses stratified by analytic approaches

	Statistics				
	Q Statistics	I^2	Multiple statistics	Other measures	Unclear
	(N = 6)	(N = 6)	(N = 2)	(N = 6)	(N = 6)
	n (%)	n (%)	n (%)	n (%)	n (%)
One-step	3 (50)	4 (67)	0 (0)	6 (100)	6 (100)
Two-step	2 (33)	2 (33)	2 (100)	0 (0)	0 (0)

This review also highlighted the strengths and weaknesses of individual patient data meta-analyses (IPD-MA) where the outcome was binary. IPD-MA are clearly the gold standard of meta-analytic methods and publications featuring results from IPD-MA are growing steadily in recent years. However, there are considerable variations in the methodology employed, for instance, the use of fixed or random effects for the estimated effect measures, measures of heterogeneity and strategies used to estimate treatment effects. In many studies, the statistical aspects were not clearly reported, with insufficient details provided to distinguish the methods used. Most times, little justification was given for the approaches taken in the studies, perhaps due to the lack of specific guidelines available for the IPD meta-analysis of binary outcomes. While guidelines exist for the reporting of systematic reviews and meta-analyses, these guidelines are not specific to IPD-MA. For example, the PRISMA guideline #14 suggests that the methods of handling data and combining results, including measures of heterogeneity be described [44]. Extending those guidelines to encompass issues specific to IPD MA, such as stating if a one- or two-stage approach was used, would likely improve the reporting of IPD meta-analyses of binary outcomes.

In a previous systematic review of articles published in 1999–2001 [13], 14 (32%) of the IPD -MA dealt with a binary outcome. While the proportion was similar, we found nearly twice the number of IPD-MA of a binary outcome in just one year in 2011.

This review of 26 IPD meta-analyses of binary outcome encouragingly shows that practitioners often obtain a large proportion of the IPD required. IPD from 90% or more of the total number of studies were obtained in 62% of IPD studies, an important improvement to the 41% found in the previous review [13].

We found that more than half (73%) of studies did not use a two-step approach (i.e. analyzing each study separately and as identically as possible and pooling via standard meta analytic methods) but instead used the more flexible one-stage method. This finding was contrary to the previous review [13], in which most analyses were performed using a two-stage approach (82%) with little consideration of the one-step approach. This finding likely reflects the greater comfort with random-effects models for binary outcomes in health research, as these models are used much more frequently now and are readily available in most mainstream statistical packages.

Heterogeneity was considered in some manner by 81% of included reviews, whether by known quantitative measures or other assessments. The most frequently used measure of heterogeneity was the I^2 statistic. Alternative measures included the Q Statistic (Chi-square statistic), and Breslow-Day test. In a few instances, heterogeneity was estimated and reported from a two-stage approach; even when a one-stage approach was used for the main analysis.

Investigating subgroup effects was one of the primary reasons for conducting an IPD-MA, and among IPD-MA that aimed to estimate a treatment effect or treatment safety all investigated subgroup effects. On the other hand, IPD-MA were unlikely to adjust for potential confounders unless observational studies were included.

Within the realm of IPD-MA with binary outcomes, our review shows that a variety of methods were used to estimate a pooled treatment effect. Many of the articles reviewed contained insufficient details on the approach used and the rationale for that approach. We next provide some recommendations and emphasize the use of the PRISMA statement to help authors ensure transparent and complete reporting of systematic reviews and meta-analyses [3,44,45]. First, if individual raw data is available for all studies and irrespective of the final approach, most statisticians and methodologists prefer the one-stage rather than a two-stage approach [2]. In some cases, the one- and two-stage approaches will give similar results [46]. However, it is currently unknown under what conditions this may be expected. Moreover, one stage methods may be preferred for evaluating treatment-covariate interactions of continuous covariates, incorporating nonlinear relationships, when studies are small, and there is heterogeneity across studies, and particularly for pooling of non randomized trials that may need to be adjusted for several confounders [46].

Moreover, methods have been developed to incorporate both individual patient data with summary level data when necessary, so that having partial IPD should not be an impediment to using a one-stage approach [5,11].

However, when random effects logistic regression is used, several details should be reported including: whether study and/or treatment were considered as random, and the statistical method used to estimate the GLMM (e.g. PQL or adaptive Gaussian Hermite quadrature). On the other hand, if a two-stage approach is used, we suggest that the meta-analytic technique used to pool results should be stated explicitly. Moreover, simply pooling subjects from various studies together is not appropriate.

Assessment and exploration of heterogeneity should always be performed in any MA, or IPD-MA. Nonetheless, how best to quantify heterogeneity remains unclear. While some advocate using the estimated variance of the random treatment effect, difficulties with its interpretation may imply that I^2 as estimated from a two-stage approach is the optimal choice for quantifying heterogeneity. Of course, whether heterogeneity estimated from a two-stage approach is relevant to a one-stage model is an open question.

There are some limitations to the work presented here. First, we have focused on binary outcomes, while survival outcomes were reported in about half of the studies retrieved (See Figure 1). Second, we limited our study retrieval to articles published in 2011. This choice was made because this gave us a sufficient sample of studies to work with that were recently completed. Moreover, we believe that there are unlikely to be major differences in the methods used, or in how they were reported between e.g. 2010 and 2011. Finally, we have focused only on the statistical approach used in these studies; whereas some may be interested more generally in how well IPD-MA are reported.

Conclusion

As found previously, we have demonstrated that a diversity of methods are employed when dealing with IPD meta-analyses for binary outcomes. Evidence from this systematic review shows that the use IPD-MA of binary outcomes has increased, with random effects logistic regression the most common method of analysis. The statistical approach taken, along with justification for that approach, is still often not reported in sufficient detail. Standardized guidelines both for the best approach to use, as well as what details to report may be needed in this area.

Additional file

Additional file 1: Table S1. Description of the 26 IPD-MA.

Abbreviations

IPD: Individual patient data; MA: Meta-analysis; IPD-MA: Individual patient data meta-analysis; IQR: Inter quartile range; PQL: Penalized Quasi-likelihood.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DT led this project in the writing and revision of the manuscript and was a reviewer. AB facilitated in the development of the scope, formulated the initial proposal and edited the final draft. SR contributed significantly in the screening and inclusion as a reviewer. All authors read and approved the final manuscript.

Acknowledgements

We thank our two reviewers Johannes Reitsma and David Fisher whose thoughtful reviews enabled us to improve our work substantially.

Author details

[1]Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montreal, Canada. [2]Department of Medicine, McGill University,

Montreal, Canada. [3]Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montreal, Canada. [4]Department of Biostatistics at the Dalla Lana School of Public Health, University of Toronto, Toronto, Canada. [5]K-129, The Montreal Chest Institute, 3650 St. Urbain, Montreal H2X 2P4, QC, Canada.

Received: 5 November 2013 Accepted: 11 June 2014
Published: 19 June 2014

References

1. DerSimonian R, Laird N: **Meta-analysis in clinical trials.** *Control Clin Trials* 1986, **7**:177–188.
2. Riley RD: **Commentary: like it and lump it? Meta-analysis using individual participant data.** *Int J Epidemiol* 2010, **39**:1359–1361.
3. Riley RD, Lambert PC, Abo-Zaid G: **Meta-analysis of individual participant data: rationale, conduct, and reporting.** *BMJ* 2010, **340**:c221.
4. Ades AE, Lu G, Higgins JP: **The interpretation of random-effects meta-analysis in decision models.** *Med Decis Making* 2005, **25**:646–654.
5. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F: **Meta-analysis of continuous outcomes combining individual patient data and aggregate data.** *Stat Med* 2008, **27**:1870–1893.
6. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI: **Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head.** *Stat Med* 2002, **21**:371–387.
7. Fisher DJ, Copas AJ, Tierney JF, Parmar MK: **A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners.** *J Clin Epidemiol* 2011, **64**:949–967.
8. Higgins JP, Thompson SG, Spiegelhalter DJ: **A re-evaluation of random-effects meta-analysis.** *J R Stat Soc Ser A Stat Soc* 2009, **172**:137–159.
9. Schmid CH, Stark PC, Berlin JA, Landais P, Lau J: **Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors.** *J Clin Epidemiol* 2004, **57**:683–697.
10. Lambert PC, Sutton AJ, Abrams KR, Jones DR: **A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis.** *J Clin Epidemiol* 2002, **55**:86–94.
11. Riley RD, Simmonds MC, Look MP: **Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods.** *J Clin Epidemiol* 2007, **60**:431–439.
12. Simmonds MC, Higgins JP: **Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data.** *Stat Med* 2007, **26**:2982–2999.
13. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG: **Meta-analysis of individual patient data from randomized trials: a review of methods used in practice.** *Clin Trials* 2005, **2**:209–217.
14. Rodseth RN, Lurati Buse GA, Bolliger D, Burkhart CS, Cuthbertson BH, Gibson SC, Mahla E, Leibowitz DW, Biccard BM: **The predictive ability of pre-operative B-type natriuretic peptide in vascular patients for major adverse cardiac events: an individual patient data meta-analysis.** *J Am Coll Cardiol* 2011, **58**:522–529.
15. van der Pas MH, Meijer S, Hoekstra OS, Riphagen II, de Vet HC, Knol DL, van Grieken NC, Meijerink WJ: **Sentinel-lymph-node procedure in colon and rectal cancer: a systematic review and meta-analysis.** *Lancet Oncol* 2011, **12**:540–550.
16. Broeze KA, Opmeer BC, Coppus SF, van GN, Alves MF, Anestad G, Bhattacharya S, Allan J, Guerra-Infante MF, Den Hartog JE, Land JA, Idahl A, Van der Linden PJ, Mouton JW, Ng EH, Van der Steeg JW, Steures P, Svenstrup HF, Tiitinen A, Toye B, Van d V, Mol BW: **Chlamydia antibody testing and diagnosing tubal pathology in subfertile women: an individual patient data meta-analysis.** *Hum Reprod Update* 2011, **17**:301–310.
17. Broeze KA, Opmeer BC, Van GN, Coppus SF, Collins JA, Den Hartog JE, Van der Linden PJ, Marianowski P, Ng EH, Van der Steeg JW, Steures P, Strandell A, Van d V, Mol BW: **Are patient characteristics associated with the accuracy of hysterosalpingography in diagnosing tubal pathology? An individual patient data meta-analysis.** *Hum Reprod Update* 2011, **17**:293–300.
18. Leroy S, Romanello C, Galetto-Lacour A, Bouissou F, Fernandez-Lopez A, Smolkin V, Gurgoze MK, Bressan S, Karavanaki K, Tuerlinckx D, Leblond P, Pecile P, Coulais Y, Cubells C, Halevy R, Aygun AD, Da DL, Stefanidis CJ, Vander BT, Bigot S, Dubos F, Gervaix A, Chalumeau M: **Procalcitonin is a predictor for high-grade vesicoureteral reflux in children: meta-analysis of individual patient data.** *J Pediatr* 2011, **159**:644–651.
19. Askie LM, Ballard RA, Cutter GR, Dani C, Elbourne D, Field D, Hascoet JM, Hibbs AM, Kinsella JP, Mercier JC, Rich W, Schreiber MD, Wongsiridej PS, Subhedar NV, Van Meurs KP, Voysey M, Barrington K, Ehrenkranz RA, Finer NN: **Inhaled nitric oxide in preterm infants: an individual-patient data meta-analysis of randomized trials.** *Pediatrics* 2011, **128**:729–739.
20. Bonati LH, Fraedrich G: **Age modifies the relative risk of stenting versus endarterectomy for symptomatic carotid stenosis–a pooled analysis of EVA-3S, SPACE and ICSS.** *Eur J Vasc Endovasc Surg* 2011, **41**:153–158.
21. von MG, Untch M, Nuesch E, Loibl S, Kaufmann M, Kummel S, Fasching PA, Eiermann W, Blohmer JU, Costa SD, Mehta K, Hilfrich J, Jackisch C, Gerber B, du BA, Huober J, Hanusch C, Konecny G, Fett W, Stickeler E, Harbeck N, Muller V, Juni P: **Impact of treatment characteristics on response of different breast cancer phenotypes: pooled analysis of the German neo-adjuvant chemotherapy trials.** *Breast Cancer Res Treat* 2011, **125**:145–156.
22. Houben RM, Crampin AC, Ndhlovu R, Sonnenberg P, Godfrey-Faussett P, Haas WH, Engelmann G, Lombard CJ, Wilkinson D, Bruchfeld J, Lockman S, Tappero J, Glynn JR: **Human immunodeficiency virus associated tuberculosis more often due to recent infection than reactivation of latent infection.** *Int J Tuberc Lung Dis* 2011, **15**:24–31.
23. Lejoyeux M, Lehert P: **Alcohol-use disorders and depression: results from individual patient data meta-analysis of the acamprosate-controlled studies.** *Alcohol Alcohol* 2011, **46**:61–67.
24. Cardwell CR, Stene LC, Joner G, Bulsara MK, Cinek O, Rosenbauer J, Ludvigsson J, Svensson J, Goldacre MJ, Waldhoer T, Jarosz-Chobot P, Gimeno SG, Chuang LM, Roberts CL, Parslow RC, Wadsworth EJ, Chetwynd A, Brigis G, Urbonaite B, Sipetic S, Schober E, Devoti G, Ionescu-Tirgoviste C, de Beaufort CE, Stoyanov D, Buschard K, Radon K, Glatthaar C, Patterson CC: **Birth order and childhood type 1 diabetes risk: a pooled analysis of 31 observational studies.** *Int J Epidemiol* 2011, **40**:363–374.
25. Patti G, Cannon CP, Murphy SA, Mega S, Pasceri V, Briguori C, Colombo A, Yun KH, Jeong MH, Kim JS, Choi D, Bozbas H, Kinoshita M, Fukuda K, Jia XW, Hara H, Cay S, Di SG: **Clinical benefit of statin pretreatment in patients undergoing percutaneous coronary intervention: a collaborative patient-level meta-analysis of 13 randomized studies.** *Circulation* 2011, **123**:1622–1632.
26. Porter CK, Riddle MS, Tribble DR, Louis BA, McKenzie R, Isidean SD, Sebeny P, Savarino SJ: **A systematic review of experimental infections with enterotoxigenic Escherichia coli (ETEC).** *Vaccine* 2011, **29**:5869–5885.
27. Groeneveld E, Broeze KA, Lambers MJ, Haapsamo M, Dirckx K, Schoot BC, Salle B, Duvan CI, Schats R, Mol BW, Hompes PG: **Is aspirin effective in women undergoing in vitro fertilization (IVF)? Results from an individual patient data meta-analysis (IPD MA).** *Hum Reprod Update* 2011, **17**:501–509.
28. Berghella V, Rafael TJ, Szychowski JM, Rust OA, Owen J: **Cerclage for short cervix on ultrasonography in women with singleton gestations and previous preterm birth: a meta-analysis.** *Obstet Gynecol* 2011, **117**:663–671.
29. de Boer SP, Barnes EH, Westerhout CM, Simes RJ, Granger CB, Kastrati A, Widimsky P, de Boer MJ, Zijlstra F, Boersma E: **High-risk patients with ST-elevation myocardial infarction derive greatest absolute benefit from primary percutaneous coronary intervention: results from the Primary Coronary Angioplasty Trialist versus thrombolysis (PCAT)-2 collaboration.** *Am Heart J* 2011, **161**:500–507.
30. Laporte S, Liotier J, Bertoletti L, Kleber FX, Pineo GF, Chapelle C, Moulin N, Mismetti P: **Individual patient data meta-analysis of enoxaparin vs. unfractionated heparin for venous thromboembolism prevention in medical patients.** *J Thromb Haemost* 2011, **9**:464–472.
31. Jefferis J, Perera R, Everitt H, Van WH, Rietveld R, Glasziou P, Rose P: **Acute infective conjunctivitis in primary care: who needs antibiotics? An individual patient data meta-analysis.** *Br J Gen Pract* 2011, **61**:e542–e548.
32. Zinkstok SM, Vergouwen MD, Engelter ST, Lyrer PA, Bonati LH, Arnold M, Mattle HP, Fischer U, Sarikaya H, Baumgartner RW, Georgiadis D, Odier C, Michel P, Putaala J, Griebe M, Wahlgren N, Ahmed N, Van GN, de Haan RJ, Nederkoorn PJ: **Safety and functional outcome of thrombolysis in dissection-related ischemic stroke: a meta-analysis of individual patient data.** *Stroke* 2011, **42**:2515–2520.
33. Saber W, Moua T, Williams EC, Verso M, Agnelli G, Couban S, Young A, De CM, Biffi R, van Rooden CJ, Huisman MV, Fagnani D, Cimminiello C, Moia M, Magagnoli M, Povoski SP, Malak SF, Lee AY: **Risk factors for catheter-**

related thrombosis (CRT) in cancer patients: a patient-level data (IPD) meta-analysis of clinical trials and prospective studies. *J Thromb Haemost* 2011, **9**:312–319.

34. Kelder JC, Cowie MR, McDonagh TA, Hardman SM, Grobbee DE, Cost B, Hoes AW: Quantifying the added value of BNP in suspected heart failure in general practice: an individual patient data meta-analysis. *Heart* 2011, **97**:959–963.
35. Waldman AT, Stull LB, Galetta SL, Balcer LJ, Liu GT: Pediatric optic neuritis and risk of multiple sclerosis: meta-analysis of observational studies. *J AAPOS* 2011, **15**:441–446.
36. Baman TS, Meier P, Romero J, Gakenheimer L, Kirkpatrick JN, Sovitch P, Oral H, Eagle KA: Safety of pacemaker reuse: a meta-analysis with implications for underserved nations. *Circ Arrhythm Electrophysiol* 2011, **4**:318–323.
37. Lanas A, McCarthy D, Voelker M, Brueckner A, Senn S, Baron JA: Short-term acetylsalicylic acid (aspirin) use for pain, fever, or colds - gastrointestinal adverse effects: a meta-analysis of randomized clinical trials. *Drugs R D* 2011, **11**:277–288.
38. He W, Gandhi CD, Quinn J, Karimi R, Prestigiacomo CJ: True aneurysms of the posterior communicating artery: a systematic review and meta-analysis of individual patient data. *World Neurosurg* 2011, **75**:64–72.
39. Black JA, Herbison GP, Lyons RA, Polinder S, Derrett S: Recovery after injury: an individual patient data meta-analysis of general health status using the EQ-5D. *J Trauma* 2011, **71**:1003–1010.
40. Breslow N, Clayton D: Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993, **88**:9–25.
41. Pinheiro J, Bates D: Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J Comput Graph Stat* 1995, **4**:12–35.
42. Higgins JP, Thompson SG: Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002, **21**:1539–1558.
43. Altman DG, Bland JM: Interaction revisited: the difference between two estimates. *BMJ* 2003, **326**:219.
44. Moher D, Liberati A, Tetzlaff J, Altman DG: Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol* 2009, **62**:1006–1012.
45. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J Clin Epidemiol* 2009, **62**:e1–e34.
46. Stewart GB, Altman DG, Askie LM, Duley L, Simmonds MC, Stewart LA: Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PLoS One* 2012, **7**:e46042.

doi:10.1186/1471-2288-14-79
Cite this article as: Thomas *et al.*: Systematic review of methods for individual patient data meta- analysis with binary outcomes. *BMC Medical Research Methodology* 2014 **14**:79.

Preamble to Manuscript II

In our published review (Manuscript I) of the methods for individual patient data meta-analysis with binary outcomes, we presented results with respect to the statistical approach used in the analysis (one- or two-step, penalized quasi-likelihood (PQL) or adaptive Gauss-Hermite quadrature (AGHQ)) and the assumption of fixed- or random-effects. Several findings were reported, however, we only considered the main conclusions of the systematic review: (i) one-step method was most frequently used (ii) estimation method (PQL or AGHQ) were not reported entirely and (iii) it was unclear if in addition to a random slope, a random intercept was also used. In the second paper (Manuscript II) of this thesis, these findings are addressed.

Manuscript II presents the next step considered in this thesis, that is, to assess via a simulation study, the performance of different analytic approaches to individual patient data meta-analyses (IPD-MA) with binary outcomes that was reported in Manuscript I. We compare (i) one-step and two-step methods, (ii) PQL and AGHQ estimation methods and (iii) stratified verses random study-effects.

Simulation comparisons between the one-step and two-step approaches to meta-analyzed IPD data are limited. Also, despite several comparisons of the estimation methods for generalized linear mixed models (GLMMs), there are few literature publications, particularly, the context of GLMMs for IPD-MA with binary outcomes (i.e. small effect size, large inter-study correlation and variance, imbalances in study sizes etc.). The lack of pragmatic guidelines and justifications has been the driving force behind this work.

The manuscript will be then submitted to an applied statistical journal such as BMC Medical Research Methodology, Journal of Statistical Computation and Simulation, or Computational Statistics & Data Analysis

Chapter 5 Manuscript II- Estimating Generalized Linear Mixed Models with Binary Outcomes for the Analysis of Individual Patient Data Meta-Analyses: which method is best?

Running title: Estimating GLMMs with Binary Outcomes for the Analysis of IPD-MA

Doneal Thomas[1], Andrea Benedetti [1, 2,3,4] and Robert Platt[1]

[1] Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montreal, Canada

[2] Department of Medicine, McGill University, Montreal, Canada

[3] Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montreal, Canada

[4] The Montreal Chest Institute, K-129, 3650 St. Urbain, Montreal, H2X 2P4, QC, Canada

Corresponding Author: Andrea Benedetti

Address: Montreal Chest Institute, K-135
3650 St. Urbain
Montreal, QC
H4A 2Z6

Telephone: (514) 934-1934 ext. 32161
Fax: (514) 843-2083
Email: andrea.benedetti@mcgill.ca

5.1 Abstract

Introduction: Individual patient data meta-analyses (IPD-MA) are regarded as the gold standard for MA and are performed using a one-step approach with increasing frequency, in a form of generalized linear mixed model (GLMM) for binary outcomes. Parameter estimation in GLMMs with binary outcomes is complicated by integrals without closed form solutions. Penalized quasi-likelihood (PQL) and adaptive Gauss-Hermite quadrature (AGHQ) methods are commonly used to circumvent this problem. Conventionally, the study effect may be modelled as either fixed or random within the GLMMs framework. If random study effects are used, the covariance between the study- and treatment-effect should be modelled and estimated.

Methods: The performance of PQL and AGHQ procedures for estimating GLMMs with binary outcomes in the context of IPD-MA were evaluated, which were then compared to the conventional approach of Der Simonian and Liard (the two-step approach) via simulation studies. The prevalence of the outcome, sample size, number of studies and variances and correlation of the random effects were varied and the comparison was done in terms of: (i) bias, (ii) mean-squared error (iii) coverage and (iv) numerical convergence, of the pooled treatment effect and inter-study heterogeneity of the treatment effect.

Results: The two-step and one-step methods produced approximately unbiased pooled treatment effect estimates, despite the advantages of the one-step in MA with 15 studies and on average 500 total subjects (small size MA). PQL for estimating the inter-study heterogeneity of the treatment effect performed better than AGHQ with respect to RMSE for small and large data sets, but absolute percent bias of the pooled treatment effect and its inter-study variability performed comparably with AGHQ for small and large data sizes. For small size MA, the random study-effects model outperformed the stratified study-effects model. However, performance was comparable for larger data sizes, but the stratified study-effect model had a slight advantage.

Conclusion: For these simulated MA, a one-step approach was recommended over the two-step method for small size MA, as it uses a more exact statistical approach and accounts for parameter correlation. Though both estimation methods can suffer from several challenges, we recommend employing the PQL procedure if interest lies in precise estimation of the inter-study heterogeneity of the treatment effect and if the major objective is the estimation of the bias of the pooled treatment effect then either estimation procedure can be applied. It should also be noted that, researchers undertaking IPD-MA with binary outcomes should always fit a random study-effect model, as it offered a more flexible fit to IPD structure and parameterizations as experienced in practice.

Keywords: Individual patient data meta-analyses; One- and two-step models; Generalized linear mixed models; Penalized quasi-likelihood; Adaptive Gauss-Hermite quadrature; Fixed and random study-effects.

5.2 Introduction

Individual Patient Data (IPD) meta-analyses (MA) are regarded as the gold standard in evidence synthesis and are increasingly being used in current practice [5 ,7]. However, some details regarding the analysis of IPD-MA remain unclear, particularly when the outcome is binary. These details include (i) should a one- or two-step model be used [17 ,20], (ii) what estimation procedure should be used to estimate the one-step model [53 ,54] and, (iii) should the study effect be fixed or random [55].

Although IPD-MA were conventionally analyzed via a two-step approach [16], over the last decade, use of the one-step approach has increased [56]. Recently, some have even suggested that the two-step and one-step framework produce similar results for MA of large randomized controlled trials [20]. The literature suggests the one-step method is particularly preferable when few studies or few events are available as it uses a more exact statistical approach than relying on a normality approximation [17].

When IPD is available and the outcome is binary, the one-step approach consists of estimating Generalized Linear Mixed Models (GLMMs) with a random slope for the exposure, to allow the exposure effect to vary across studies. Penalized quasi-likelihood (PQL) introduced by Breslow and Clayton is the most method popular for estimating the parameters in GLMMs [35]. However, regression parameters can be badly biased for some GLMMs, especially with binary outcomes with few observations per cluster, low outcome rates, or high between cluster variability [38 ,41].

Adaptive Gaussian Hermite quadrature (AGHQ) is the current favored competitor to PQL, which approximates the likelihood by numerical integration [36]. Although estimation becomes more precise as the number of quadrature points increases, it often gives rise to computational difficulties for high-dimension random effects and convergence problems where variances are close to zero or cluster sizes are small [36].

The heterogeneity between studies is an important aspect to consider when carrying out IPD-MA. Such heterogeneity may arise due to differences in study design, treatment protocols or patient populations [55]. When such heterogeneity is present, the convention is to include a random slope in the model as it captures the variability of the exposure across studies. However, there is a considerable amount of controversy in regards to the study effect being modelled as stratified or random [34]. When a study is considered as random, it is assumed hat the log-odds are drawn from a normal distribution [34], while the stratified study effect estimates a separate intercept for each included study, in the absence of the normality assumption.

Few comparisons have been reported in the context of GLMMs for IPD-MA with binary outcomes [17 ,34] - when the number of study and the number of subjects within the study is small, imbalanced study sizes, large between-study heterogeneity, small exposure effects and an interest in the variance parameter of the treatment effect. According to previous literature, these factors have all been identified as influencing model performance [53]. In addition, several simulation studies limit their attention to simple models with only random intercepts, as a results, the performance of the random effects models including both a random intercept and a random slope are far less reported.

Our objective was to assess and compare via simulation studies, (i) the performance of different estimation procedures for GLMMs with binary outcomes, (ii) compare using stratified study-effect and random study-effects estimates in one-step approaches to two-step methods. Moreover, we used these findings to develop guidelines on the choice of methods for analyzing data from IPD-MA with binary outcomes and to understand explicitly the trade-offs between computational and statistical complexity.

Section 5.3 described briefly the GLMM and the estimation procedures. Section 5.4 introduced the models we are considering, the design of the simulation study and the assessment criteria. In section 5.5, results for the different methods under varying conditions are presented and discussed. Section 5.6 concluded with a discussion.

5.3 Generalized linear mixed models and estimation methods

The extension of the generalized linear models with random effects terms is called the GLMM [31]. The conditional independence assumption of the outcome, given the random effects in a GLMM, is essential in the formulation of the joint likelihood function. However, the main difficulty with GLMM estimation is that no closed analytic solutions for the joint likelihood function are available but a number of effective ways to compute and maximize the likelihood have been developed. In this paper, we considered the two mainstream techniques: (i) PQL [35]; and (ii) AGHQ [36].

5.4 Methods

We conducted a simulation study to compare (i) one- vs. two-step approaches, (ii) PQL vs. AGHQ and (iii) random- vs. stratified study-effect, when analyzing data from IPD-MA with binary outcomes. Hereto, our methods all assume that between-study heterogeneity exists, as it is likely in practice, and so only random treatment-effects IPD meta-analysis models are considered.

5.4.1 Data Generation

The data generation algorithm was developed to generate two-level data sets (e.g. patients grouped into studies). We generated a binary outcome (Y_{ij}) and a single binary exposure (X_{ij}). We denote the number of studies $j = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, n_j$ denotes the individuals per study. Therefore, Y_{ij} is the outcome observed for the i^{th} individual from the j^{th} study.

The dichotomous exposure variable, X_{ij}, was generated from a Bernoulli distribution with probability = 0.5 and recoded $\pm$½ to indicate control/treatment group. The coding of $\pm$½ is advantageous when fitting a random effects meta-analysis model with random study effects in data sets with few degrees of freedom, and where estimation of a covariance between two random effects is problematic or impossible [34]. To generate the binary outcome variable Y_{ij}, first the probability of the outcome was calculated from the random-study and –treatment effects logistic regression model given by Equation (1), and the stratified-study effects model given by Equation (2):

$$logit(\pi_{ij}) = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j})z_{ij} \cdots (1)$$

$$logit(\pi_{ij}) = \beta_k + (\beta_1 + b_{1j})z_{ij} \cdots\cdots\cdots (2)$$

Here π_{ij} is the true probability of the outcome for the i^{th} individual from the j^{th} study, β_0 denotes the mean log-odds of the outcome and β_1 the pooled treatment effect (log odds ratio). The random effects (b_{0j} and b_{1j}) were generated from a bivariate normal distribution with mean = 0 and variance-covariance matrix $\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}$ for the random study-effect case. In the stratified study effects case, (i.e. Eq. (2)), β_k, were generated from a uniform distribution and b_{1j} was generated from a normal distribution with zero mean and variance, τ^2.

A Bernoulli distribution with probability π_{ij} from Equations (1) and (2) was used to generated the binary outcome Y_{ij}.

The number of studies, study size, total sample size, variances and correlation of the random effects, and average conditional probability were all varied, with levels described in Table 1. For each distinct combination (n=96) of simulation parameters, 1000 IPD-MA were generated from each equation (1) and (2), allowing us to investigate a wide range of scenarios.

Table 1[a]: Summary of Simulation Parameters

Parameters	Values
IPD-Meta-analyses generated:	M=1000
(Number of studies, number of subjects per study, total average sample sizes)[b]:	$(K, n_i, N) \in$ {(5,100,500), (15,33,500), (15,200,3000), **(5,357,500)**, **(15,98,500)**, **(15,588,3000)**}
Fixed effects (intercepts):	$\beta_0 = -0.85$
Prevalence of the outcome	$\pi = 30\%$
Fixed effects (Slopes):	$\beta_1 = 0.18$
Random effects distribution:	Normal
Random effects variances:	$\{\sigma^2, \tau^2\} \in (0.05, 1, 4)$
Correlation between random effects:	$\rho \in (0, 0.5)$

[a] In a sensitivity analysis, we extended the number of studies to 50 with an average sample size of 9000 and reduce the prevalence of the outcome to 5%. The prevalence of the outcome was fix to 30% by setting the value of the intercept β_0 to -0.85. [b] The number of subjects per study was reported for only large studies when data sets were generated with imbalance study sizes (**bold text**: 25% large studies-10 times more subjects thank normal size).

A sensitivity analysis was also considered to explore the performance of different methods to a 5% outcome rate. All simulation parameters were held constant, however, the number of studies and the distribution of study sizes were allowed to vary.

5.4.2 Models

1. Two-step IPD methods

In the two-step approach, each study in the IPD was analyzed separately via logistic regression

$$y_{ij} \sim Bernoulli(\pi_{ij})$$

$$logit(\pi_{ij}) = \beta_0 + \beta_1 z_{ij}$$

The first step estimates the intercept, the slope and their associated within-study covariance matrix (consisting the variances of the intercept and slope, as well as the covariance) for each study. This model reduces the IPD to AD for each study. At the second stage the effect estimates are synthesized.

Model 1- Bivariate meta-analysis:

The AD (here intercept and slope) are combined via a bivariate random-effects model that simultaneously synthesizes the estimates whilst accounting for their correlation [17]. The model assumes that the true effects follow a bivariate normal distribution and is estimated via restricted maximum likelihood with the following marginal distributions of the estimates [57]:

$$\begin{bmatrix}\widehat{\beta_{0J}}\\ \widehat{\beta_{1J}}\end{bmatrix} \sim N\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \Sigma + \mathrm{C}_j\right), \Sigma = \begin{pmatrix}\sigma^2 & \rho\sigma\tau\\ \rho\sigma\tau & \tau^2\end{pmatrix}$$

where Σ is the unknown variance-covariance matrix of the true effects (β_0 and β_1) and C_j $(j = 1, \ldots, K)$ the diagonal matrix with the variances of the estimates.

Model 2: Conventional DerSimonian and Laird approach [1]:

The with-in study and between-study covariance estimates are often times ignored, and instead a univariate meta-analysis of the logit of the odds ratios is performed [58]. The marginal distribution of the pooled estimated treatment effect under this approach is easily obtained as:

$$\widehat{\beta_{1J}} \sim N(\beta_1, \tau^2 + \sigma_j^2)$$

with unknown parameters β_1 and τ^2, estimated via the inverse variance weighted non-iterative method (method-of-moment).

2. One-step methods

The one-step approach analyzes the IPD from all studies simultaneously, while accounting for clustering of subjects within studies [17]. The one-step model is a form of GLMM. Two different specifications are considered.

Model 3- Random intercept and random slope

We estimated a GLMM with a random study effect b_{0j} and a random treatment effect b_{1j} via PQL and AGHQ, and allowed the random effects to be correlated. Note that the covariance between the b_{0j} and the b_{1j} may be estimated.

$$logit(\pi_{ij}) = \beta_0 + b_{0j} + (\beta_1 + b_{1j})z_{ij}$$

$$\begin{bmatrix} b_{0j} \\ b_{1j} \end{bmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_j\right), \Sigma_j = \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}$$

Model 4-Stratified intercept one-step

Finally, the stratified one-step approach estimates a separate intercept for each study rather than constraining the intercepts to follow a normal or other distribution. Therefore, there is no need for the normality assumption for the study membership, hence, the between-study covariance term is no longer estimated. The model is defined as follows:

$$logit(\pi_{ij}) = \sum_{k=1}^{K} (\beta_k I_{k=j}) + (\beta_1 + b_{1j})z_{ij}$$

where $I_{k=j}$ indicates that a separate intercept should be estimated for each study $j = 1, \ldots, K$ and $b_{1j} \sim N(0, \tau^2)$. These models were estimated via PQL and AGHQ.

5.4.3 Estimation Procedures and Approximations

The parameters of the one-step models were estimated using PQL, AGHQ, while that of the two-step model were estimated via method-of-moments (MOM) (Model 2) and restricted maximum likelihood (REML) (Model 1) [1 ,59 ,60] at the second stage.

Both likelihood-based methods (PQL and AGHQ) were implemented on SAS version 9.4 using PROC GLIMMIX [61]. REML estimation was chosen for PQL and maximum likelihood (ML) for AGHQ.

Therefore, for each generated data set the following 4 models were fit.

- Two-step approach (Models 1 and 2)
- One-step approach via GLMMs (Models 3 and 4) estimated with PQL and AGHQ.

5.4.4 Assessment criteria

The performance of the estimation methods was evaluated using: a) numerical convergence, b) absolute bias; c) root mean square error (RMSE); and d) coverage probability - of the pooled treatment effect and its inter-study variability.

Numerical convergence

The convergence rate was estimated for both PQL and AGQH procedures, as the number of simulation repetitions that did converge (without returning a warning message) divided by the total attempted (M=1000). This was iteratively decided using the relative deviation of the parameter estimates [62].

Bias

The Monte Carlo bias of the pooled treatment effect and its inter-study heterogeneity is defined as the average of the bias in the estimates provided by each method as compared to the truth, across the 1000 IPD-MA in each scenario. The Monte Carlo estimate of the bias is computed as

$$bias = \frac{1}{1000}\sum_{j=1}^{1000} \hat{\theta}_j - \theta$$

Positive and negative bias represents over- and under-estimation for each method respectively, but does not provide an overall measure of bias. Therefore, we also reported the mean absolute bias (AB) via the following formula:

$$AB = \frac{1}{1000}\sum_{j=1}^{1000} \left|\hat{\theta}_j - \theta\right|,$$

where $\hat{\theta}_j$ were the parameter estimates and θ was the true parameter of the pooled treatment effect or its inter-study variance.

Mean square error

The mean square error (MSE) is a useful measure of the overall accuracy, because it penalizes an estimate for both bias and inefficiency. The Monte Carlo estimate of the MSE is:

$$MSE(\hat{\theta}) = \frac{1}{1000}\sum_{j=1}^{1000} \widehat{(\theta}_j - \theta)^2,$$

For each scenario, the RMSE of the pooled treatment effect and its inter-study heterogeneity is reported, as this measure is on the same scale as the parameter.

Coverage probability

We estimated coverage for the pooled treatment effect and its inter-study heterogeneity for the various methods. Gaussian coverage was estimated, where if $\left|\hat{\theta} - \theta\right| \leq 1.96 \times SE(\hat{\theta})$ the true value was covered, and if $\left|\hat{\theta} - \theta\right| > 1.96 \times SE(\hat{\theta})$ it was not.

We reported the median, the 25th and 75th percentiles of the AB and RMSE of the pooled treatment effect and its inter-study heterogeneity but reported percentages for the numerical convergence and coverage rate.

5.5 Results

Tables 2 and 3 present the median and interquartile range of the AB, RMSE, coverage and convergence of the pooled treatment effect and its inter-study variance, respectively, as estimated via two- and one-stage; AGHQ and PQL; random slope only and random intercept and slope methods. We reported results for data generated with imbalances in study sizes (different sample size in all studies) and random intercept data generation (equation 1) with correlated random effects, as this scenario is likely the closest to real-life.

We excluded results from meta-analyses that failed to converge.

Table 2 Performance of the different approaches in small data sets[a] with greater (Top panel) and lesser (Bottom panel) heterogeneity of random effects.

		Methods[b]					
	Performance measures[c]	Two-step vs.	One-step	AGHQ vs.	PQL	Random Slope only vs.	Random intercept & Slope
(τ^2, σ^2) = (4,4)[d]	AB (β_1)	0.04 (0.02 0.06)	0.04 (0.02, 0.07)	0.05 (0.02, 0.08)	**0.04** **(0.02, 0.07)**	0.04 (0.02, 0.08)	0.04 (0.02, 0.07)
	RMSE (β_1)	**1.11** **(0.49, 1.94)**	1.19 (0.53, 2.12)	1.42 (0.64, 2.52)	**1.19** **(0.53, 2.12)**	1.24 (0.49, 2.44)	**1.19** **(0.53, 2.12)**
	Coverage (β_1)	89.3	**91.8**	**93.2**	91.8	**99.1**	91.8
	AB (τ^2)	0.23 (0.14,0.30)	**0.16** **(0.08, 0.24)**	0.18 (0.09,0.29)	**0.16** **(0.08, 0.24)**	0.16 (0.07, 0.25)	0.16 (0.08, 0.24)
	RMSE (τ^2)	7.26 (4.39,7.51)	**4.93** **(2.56, 7.51)**	5.76 (2.80,9.07)	**4.93** **(2.56, 7.51)**	5.01 (2.35,7.95)	**4.93** **(2.56, 7.51)**
	Coverage (τ^2)[e]	NA	NA	**85.5**	76.2	11.6	**76.2**
	Convergence	**100**	97.7	**99**	97.7	13.8	**97.7**
(τ^2, σ^2) = (1,1)	AB (β_1)	0.02 (0.01, 0.04)	0.02 (0.01, 0.04)	0.03 (0.01, 0.05)	**0.02** **(0.01, 0.04)**	0.03 (0.01, 0.04)	**0.02** **(0.01, 0.04)**
	RMSE (β_1)	**0.73** **(0.35, 1.29)**	0.75 (0.37, 1.33)	0.79 (0.41, 1.42)	**0.75** **(0.37, 1.33)**	0.83 (0.41, 1.38)	**0.75** **(0.37, 1.33)**
	Coverage (β_1)	89.1	**90.6**	**92.3**	90.6	**96.4**	90.6
	AB (τ^2)	0.06 (0.03, 0.08)	**0.04** **(0.02, 0.1)**	0.06 (0.03, 0.09)	**0.04** **(0.02, 0.1)**	0.05 (0.03, 0.09)	**0.04** **(0.02, 0.1)**
	RMSE (τ^2)	1.73 (0.85, 2.65)	**1.22** **(0.53, 3.16)**	1.76 (0.84, 2.70)	**1.22** **(0.53, 3.16)**	1.72 (0.85, 2.78)	**1.22** **(0.53, 3.16)**
	Coverage (τ^2)	NA	NA	74.5	**81.1**	37.3	**81.1**
	Convergence	**100**	90.4	**96.8**	90.4	42.6	**90.4**

[a] Small data sets had 15 studies and on average 500 total subjects.

[b] Estimation by PQL was used for brevity (most commonly used) to compare one-step versus two-step, as well as random slope only model versus random intercept and slope method. Only the Der Simonian and Laird two-step method (Model 2) was used for comparison (conventional method). Superior measures were highlighted in bold text.

[c] Median (25th and 75th percentile) were reported for AB and RMSE, the proportion was reported for coverage and convergence.

[d] (τ^2, σ^2)**: (**Random treatment-effect variance, random study-effect variance**).**

[e] The two-step approach did not return a confidence interval for τ^2, hence no coverage was estimated and comparison was not applicable (NA) to the one-step method.

Table 3 Performance of the different approaches in large data sets[f] with greater (Top panel) and lesser (Bottom panel) heterogeneity of random effects.

		Methods[g]					
	Performance measures[h]	Two-step vs.	One-step	AGHQ vs.	PQL	Random Slope only vs.	Random intercept & Slope
(τ^2, σ^2) = (4,4)[i]	AB (β_1)	0.03 (0.02 0.06)	0.03 (0.02, 0.06)	0.04 (0.02, 0.06)	**0.03** **(0.02, 0.06)**	0.04 (0.02, 0.06)	**0.03** **(0.02, 0.06)**
	RMSE (β_1)	**1.02** **(0.50, 1.85)**	1.07 (0.49, 1.84)	1.20 (0.55, 1.99)	**1.07** **(0.49, 1.84)**	1.11 (0.55, 1.94)	**1.07** **(0.49, 1.84)**
	Coverage (β_1)	91.9	**92.3**	92.2	**92.3**	**95.2**	92.3
	AB (τ^2)	0.14 (0.07,0.22)	**0.12** **(0.06, 0.20)**	0.13 (0.07,0.21)	**0.12** **(0.06, 0.20)**	0.13 (0.06, 0.20)	**0.12** **(0.06, 0.20)**
	RMSE (τ^2)	4.36 (2.22,6.80)	**3.87** **(1.81, 6.21)**	4.12 (2.06,6.77)	**3.87** **(1.81, 6.21)**	4.05 (1.85,6.25)	**3.87** **(1.81, 6.21)**
	Coverage (τ^2)[j]	NA	NA	**81.9**	78.9	53.7	**78.9**
	Convergence	**100**	98.3	**100**	98.3	63.8	**98.3**
(τ^2, σ^2) = (1,1)	AB (β_1)	0.02 (0.01, 0.03)	0.02 (0.01, 0.03)	0.02 (0.01, 0.03)	0.02 (0.01, 0.03)	0.02 (0.01, 0.03)	0.02 (0.01, 0.03)
	RMSE (β_1)	0.61 (0.30, 1.04)	**0.59** **(0.29, 1.05)**	0.60 (0.30, 1.06)	**0.59** **(0.29, 1.05)**	0.63 (0.29, 1.07)	**0.59** **(0.29, 1.05)**
	Coverage (β_1)	91.2	**91.9**	91.7	**91.9**	91.8	**91.9**
	AB (τ^2)	0.03 (0.02, 0.06)	0.03 (0.02, 0.05)	0.04 (0.02, 0.06)	**0.03** **(0.02, 0.05)**	0.03 (0.02, 0.06)	0.03 (0.02, 0.05)
	RMSE (τ^2)	1.08 (0.53, 1.73)	**1.03** **(0.49, 1.68)**	1.09 (0.52, 1.75)	**1.03** **(0.49, 1.68)**	1.06 (0.52, 1.74)	**1.03** **(0.49, 1.68)**
	Coverage (τ^2)	NA	NA	**83.6**	82.5	**86.3**	82.5
	Convergence	**100**	96.5	**99.5**	96.5	95.3	**96.5**

[f] Large data sets had 15 studies and on average 3000 total subjects.

[g] Estimation by PQL was used for brevity (most commonly used) to compare one-step versus two-step, as well as random slope only model versus random intercept and slope method. Only the Der Simonian and Laird two-step method (Model 2) was used for comparison (conventional method). Superior measures were highlighted in bold text.

[h] Median (25th and 75th percentile) were reported for AB and RMSE, the proportion was reported for coverage and convergence.

[i] (τ^2, σ^2)**: (**Random treatment-effect variance, random study-effect variance**).**

[j] The two-step approach did not return a confidence interval for τ^2, hence no coverage was estimated and comparison was not applicable (NA) to the one-step method.

5.5.1 One- versus Two-step

- Absolute bias (AB) of the pooled treatment effect, β_1 estimates was similar and under 0.05 in the one step and the two-step approaches in both small and large data sets (Table 2 and 3).
    - The difference in the amount of bias depended on the true τ^2, and the number of studies.
    - Absolute percent bias in β_1 was reduced when data was generated to have balanced study sizes (same sample size in all studies) and when the intercept was generated as stratified (results not shown)- less variation between studies.
    - Both the one- and the two-step methods continued to produce negligible and similar bias for β_1, when the outcome rate was reduced from 30% to 5% (Figure 1 in the Appendix A).
- Root mean square error (RMSE) in β_1 was slightly larger when the one-step method was used, and the total sample size was large or small (Table 2 and 3).
    - Precision in the estimates of β_1 increased when data was generated with balanced study sizes (results not shown) and the true heterogeneity in the random effects was reduced (Bottom panel of Table 2 and 3).
    - RMSE in β_1 was inflated when the outcome rate was reduced for both methods, and the two-step approach continued to yield lower RMSE in β_1 estimates (Figure 2 in the Appendix A).
- Percent coverage of β_1 was usually under nominal levels for the both approaches and somewhat higher for the one step approach (Table 2 and 3). Both approaches did better with increased total sample size (Table 3).
    - Percent coverage of β_1 increased when data was generated with equal study sizes (results not shown) and, decreased when the outcome rate was reduced, particularly when true heterogeneity was large (Table 3 in the Appendix A). However, the two-step method still yielded percent coverage under nominal level.
- Absolute bias of the inter-study heterogeneity, τ^2 was usually slightly lower when the one step approach was used than when the two-step approach was, particularly, when the

sample size was small and when greater amount of heterogeneity exist in the random effects (Bottom panel of Table 2).

  - The amount of bias decreased when data was generated with equal study sizes (results not shown) and when total sample size increased (Table 3). The bias increased when the rate of occurrence was reduced. In these cases, the one-step approach was most biased (Figure 3 in the Appendix A).

- RMSE of $\tau^2$was mostly lower when the one step approach was used, though the difference was less as total sample size increased and when less heterogeneity was in the random effects (Bottom panel of Table 3).
  - RMSE of τ^2 was still lower when the one-step method was used and the outcome rate was reduced (Figure 4 in the Appendix A).
- Convergence was consistently higher for the two-step approach than for the one-step approach (Table 2 and 3).

5.5.2 AGHQ versus PQL

- Absolute bias in β_1 was usually similar, but slightly greater when AGHQ was used than when PQL was used (Table 2 and 3).
  - The bias was reduced when total sample size was increased (Table 3)
  - Similar amount of bias in β_1 was observed when the outcome rate was reduced (Figure 1 in the Appendix A).
- RMSE estimates of β_1 were generally greater when AGHQ was used than when PQL was used (Table 2 and 3).
  - When total sample size was increased (Table 3) or when data was generated with equal study sizes (results not shown), the RMSE was significantly reduced.
  - RMSE of β_1 was inflated when the event rate was reduced (Figure 3 in the Appendix A).

- Coverage for β_1 was always closer to nominal levels with AGHQ than with PQL, particularly, when the true heterogeneity was large and total sample was small (Top panel of Table 2).

    - Coverage was closer to nominal levels for PQL when the number of studies and total sample size was larger (Table 3).
    - Similar coverage of the estimates was observed when the outcome rate was reduced (Table 3 in the Appendix A).
- Absolute bias in τ^2 was very high, and similar when PQL and AGHQ were used, when the number of studies was small and the true heterogeneity was large (Top panel of Table 2).
    - When the total average sample size was increased, the bias was lower when PQL was used (Table 3). The difference was smaller when data was generated with a stratified intercept (data not shown).
    - Even greater bias in τ^2 estimates was observed when the event rate was reduced (Figure 3 in the Appendix A).
- RMSE of τ^2 was generally lower with PQL than with AGHQ (Table 2 and 3).
    - PQL-estimates continued to yield smaller RMSE than AGHQ-estimates when the outcome rate was reduced (Figure 4 in the Appendix A).
- Percent coverage of τ^2 was greatly under-covered for both estimation methods, particularly when PQL was used. (Table 2 and 3). This pattern was also evident when the outcome rate was reduced (Table 5 in the Appendix A).
- Convergence occurred more often when AGHQ was used than when PQL was used (Table 2 and 3).
    - Convergence was problematic for PQL, particularly when true heterogeneity was low and number of studies was few (Bottom panel of Table 2).
    - Similar convergence was seen when the event rate was reduced (Table 4 in the Appendix A).

5.5.3 Random intercept & slope versus random slope only

- Absolute bias in β_1 was similar for both random slope only and random intercept & slope methods when PQL was the choice of estimation than when AGHQ was used (Table 2 and 3)

    - The difference was negligible when a random slope and intercept was fit than when a random slope only was fit.
    - Negligible and comparable bias for β_1 was observed when the event rate was reduced (Figure 1 in the Appendix A).

- RMSE in β_1 was smaller when estimated by the random intercept and slope model than when only a random slope was fit and when the number of studies was small and large (Table 2 and 3).
    - RMSE in β_1 was still smaller when the random intercept and slope was fit and the outcome rate was reduced (Figure 2 in the Appendix A).
- Coverage of β_1 was conservative, when a random slope only was fit in small sample than when fit with a random slope and intercept (Table 2).
- Absolute bias in τ^2 was similar when fit with a random intercept & slope approach or a random slope only (Table 2 and 3).
    - There was a trend towards less bias when a random slope only was fit and when the outcome rate was reduced (Figure 3 in the Appendix A).
- RMSE of τ^2 was mostly lower when a random intercept was fit, especially when the true heterogeneity was large (Top panels of Table 2 and 3).
    - RMSE of τ^2 was comparable when both models were fit in large sample and the true heterogeneity was small (Bottom panel of Table 3)- also when outcome rate was reduced (Figure 4 in the Appendix A).
- Percent coverage of τ^2 was greatly under-covered when both models were fit and the true heterogeneity was large, particularly, when a random slope only model was fit (Top panels of Table 2 and 3).
    - Coverage continued to be an issue when equal study sizes were used (data not shown); when the rate of occurrence was reduced (Table 5 in the Appendix A).
- Convergence was markedly different for both a random intercept & slope method, and a random slope only method when fit in small data sets (Table 2).
    - In these cases, the random intercept & slope method convergence more often than the random slope only approach.
    - Convergence improvement when the total sample size increased (Table 3). The

random intercept & slope approach continued to converge often than the random slope only approach for thee data.

5.6 Discussion

Findings

Our simulation results indicate that when the number of subjects per study is large, the one- and two-step methods yield very similar results. Our results also confirm the finding of previous empirical studies [20 ,63 ,64] that in some cases, the one-step and two-step IPD-MA results coincide. However, we found discrepancies between these methods, with a slight preference towards the one-step method when the number of subjects per study is small. In these situations, neither method produced accurate estimates for the inter-study heterogeneity associated with the treatment-effect; however, the biases were larger for the two-step approach. Furthermore, one-step methods produced less biased and more precise estimates of the variance parameter and had slightly higher coverage probabilities.

Estimation of GLMMs with binary outcomes continues to pose challenges, with many methods producing biased regression coefficients and variance components [54]. AGHQ has been shown to overestimate the variance component with few clusters or few subjects [65]. On the contrary, PQL has been found to underestimate the variance component while the standard errors are overestimated [41]. We found that the absolute bias of the PQL-estimated pooled treatment effect was slightly less than that of the AGHQ-estimates. The PQL-estimates of the inter-study variance had greater precision when study sizes were small and random effects were correlated. This somewhat confirms previous results, which found that PQL suffers from large biases but performs better in terms of MSE than AGHQ [53]. Both estimation methods experienced difficulty in attaining nominal coverage of the inter-study heterogeneity associated with the treatment effect in two situations: (i) when the number of studies included was small and/or (ii) the true variances between random effects was small. AGHQ-estimated confidence intervals around the inter-study heterogeneity of the pooled treatment effect were greatly under nominal levels in these scenarios. For PQL-estimated confidence interval, at times severe under coverage for the inter-study heterogeneity was found, both because the PQL estimates were downward biased and the standard errors were too small. Our work also found that convergence was not a significant problem for AGHQ when meta-analyses include study sizes with less than 50 individuals per study. The problem was exacerbated when the prevalence of the outcome was reduced to 5% and the true heterogeneity was close to zero.

Stratification of the intercept in the one-step models avoids the need to estimate the random effect for the intercept and the correlation between the random effects. Overall, we found that the random slope only approach suffered from marked convergence rates when fit to small data sets (15 studies and on average 500 subjects). It was also found that the loss in efficiency was due to fitting several dummy variables for the study effect rather than a normality assumption. V. Rondeau et al. also found that with survival multilevel data, not including the intercept as random in the additive Cox model could lead to inaccurate results [55]. We found that the absolute percent bias of the treatment effect and its associated variance was comparable for both models. The random intercept and slope model was found to produce the most precise estimates of the pooled treatment effect and its inter-study heterogeneity in small data sets. For the random slope only model, the coverage rate was close to and above nominal levels for the estimates for the pooled treatment effect, but suffered from convergence, as the random slope only model might be under-defined (lack of sufficient variation in the intercept). We found that both models failed to achieve close to nominal coverage rates of the inter-study heterogeneity of the treatment effect, as the construction of the confidence interval are likely to be invalid [34].

Strengths and Limitations

We used simulation studies to compare various analytic strategies to analyze data arising from IPD-MA across a wide range of data generation scenarios but made some simplifications. We only considered binary outcomes, one dichotomous treatment variable, a two-level data structure, and no confounders-however; this is the least common reason provided to support conducting IPD-MA [56]. Moreover, we estimated GLMMs via PQL and AGHQ, but did not compare Bayesian or other estimation methods, which might be particularly useful in sparse scenarios [66]. Throughout this thesis, we have assumed that IPD were available. Certainly, the time and cost associated with collecting IPD are considerable. However, once such data is in hand, we have addressed several open questions relating to the best way to analyze it. We should also note that methods exist for combing IPD and aggregated data [7]. Further study is needed to investigate alternative confidence intervals (or coverage probability) for the inter-study heterogeneity that can be used to remedy the under-coverage of Gaussian intervals. The normality based intervals (coverage rate) we have studied, greatly underperformed in most scenarios. A further simplification that limits the generalizability of this work is that it is restricted to only two-arm trials. The extension to three or more arms would require careful consideration of more complicated correlation structures in treatment effects across arms and within studies [67].

One important comparison we have not addressed is, computational speed where the two-step method had a distinct advantage over the one-step; PQL was faster than AGHQ and the stratified-intercept model run-time was quicker than the random-intercept model.

As far as we know, this simulation study is the first to simultaneously generate data with normally distributed and stratified random intercepts. This study also compares approaches that include a random intercept for study membership to those that do not. Furthermore, the use of simulation - to systematically investigate the robustness of the approaches to variation in sample size, study number, outcome rate, magnitude of correlation and variances. As a result, our scenarios have allowed us to assess performance without being too exhaustive.

Guidelines for Best Practice

On the basis of these findings, we can make several recommendations. When the IPD-MA included many studies and the outcome rate was not too low, this work supports the conclusion of a previous study [20] that the conventional two-step method by DerSimonian and Laird [1]is a good choice. Further, while the bivariate two-step approach is very rarely used in practice, we found that it tended to yield good overall model performance, comparable with that of the one-stage models when study sizes are small. In addition, our results also suggest that the one-step method can be used in IPD-MA where study sizes are less than 50 subjects per study or few events were recorded in most studies (outcome rate of 5%). In these cases, the one-step approach is more appropriate as it models the exact binomial distribution of the data and offers more flexibility in model specification.

If interest lies in estimation of the pooled treatment effect or the inter-study heterogeneity of the treatment effect, estimation using PQL appeared to be a better choice due to its excellent coverage probabilities and lower bias for the settings considered. In addition, computational issues such as convergence occurred less with this technique than with AGHQ. Finally, if one were interested in precise estimation of the inter-study heterogeneity of the treatment effect, the PQL approach may be preferred. However, it is important to note that convergence and coverage in τ^2 was an issue in small and large total sample sizes and also, when level of true heterogeneity was too large.

Fitting GLMMs with random intercept and random slope were more robust than stratification of the intercept when conducting IPD-MA with binary responses. These models always estimate the variability between the random effects from the data without any assumptions.

There are four important caveats to these recommendations. First, our simulations show greater accuracy of the pooled odds ratio as the number of studies increase. Therefore, an IPD-MA with more studies will provide more accurate estimates. Secondly, our results show that the estimation of the inter-study heterogeneity of the treatment effect is highly biased regardless of the sample size and number of studies. Therefore, we should always expect that the variance parameter be estimated with some error. Thirdly, small overall samples mark the trade-off under which a meta-analyst might consistently choose precision over bias and our simulations show that AGHQ estimation may be preferred in these situations. Finally, large overall sample size can eliminate the lack of statistical power present in small overall samples. In such cases, comparable results are seen for one- and two-step methods and fitting a two-step analysis as a first step may be advisable. This could aid as a quick and efficient investigation of heterogeneity and treatment-outcome association.

Conclusion

To summarize, the one- and two-step methods consistently produced similar results when the number of studies and overall sample are large. Although the PQL and AGHQ estimation procedures produced similar bias of the pooled log odds ratios, PQL-estimates had lower RMSE than the AGHQ-estimates. The random intercept and slope model yielded precise estimates and good coverage probabilities of the pooled treatment effect and its inter-study heterogeneity in small and large overall sample sizes as compared to the random slope only model.

Authors' contributions

DT led this project in the study design, performed simulation of data and statistical analyses, and also led the writing of the manuscripts. AB participated in the study design, guided statistical analyses and edited the final draft. RP helped draft and revised the manuscript. All authors read and approved the final manuscript.

Appendix A: Sensitivity analysis results

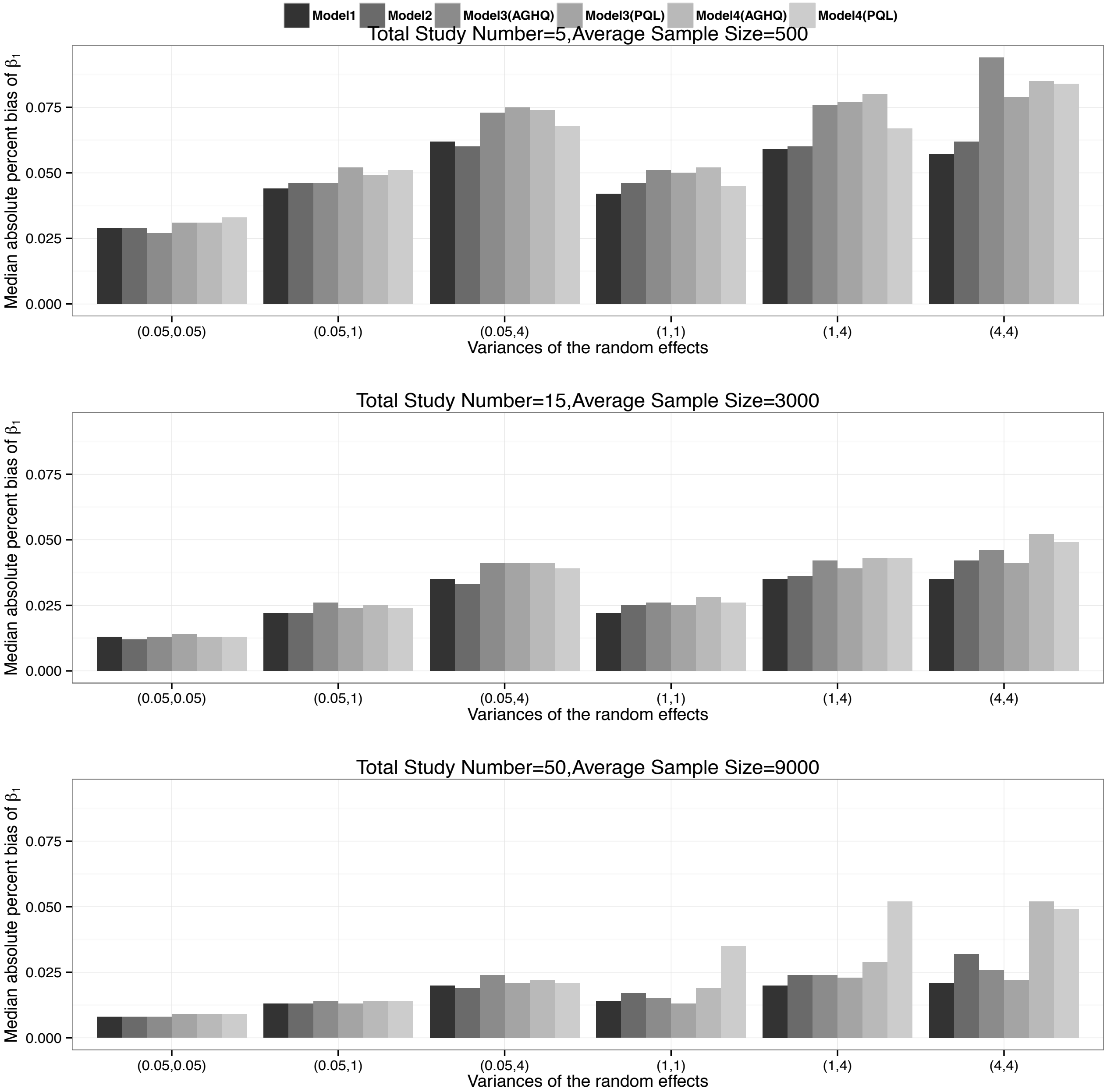

Figure 1 Median absolute bias of the pooled treatment effect, β_1, by number of studies, estimation methods and sample sizes over increasing random-effect variances: Bivariate two-step (Model1), DeSermonian-Laird two-step (Model2), random intercept one-step (Model3; estimation method: PQL or AGHQ) and stratified intercept one-step (Model4; estimation method: PQL or AGHQ).

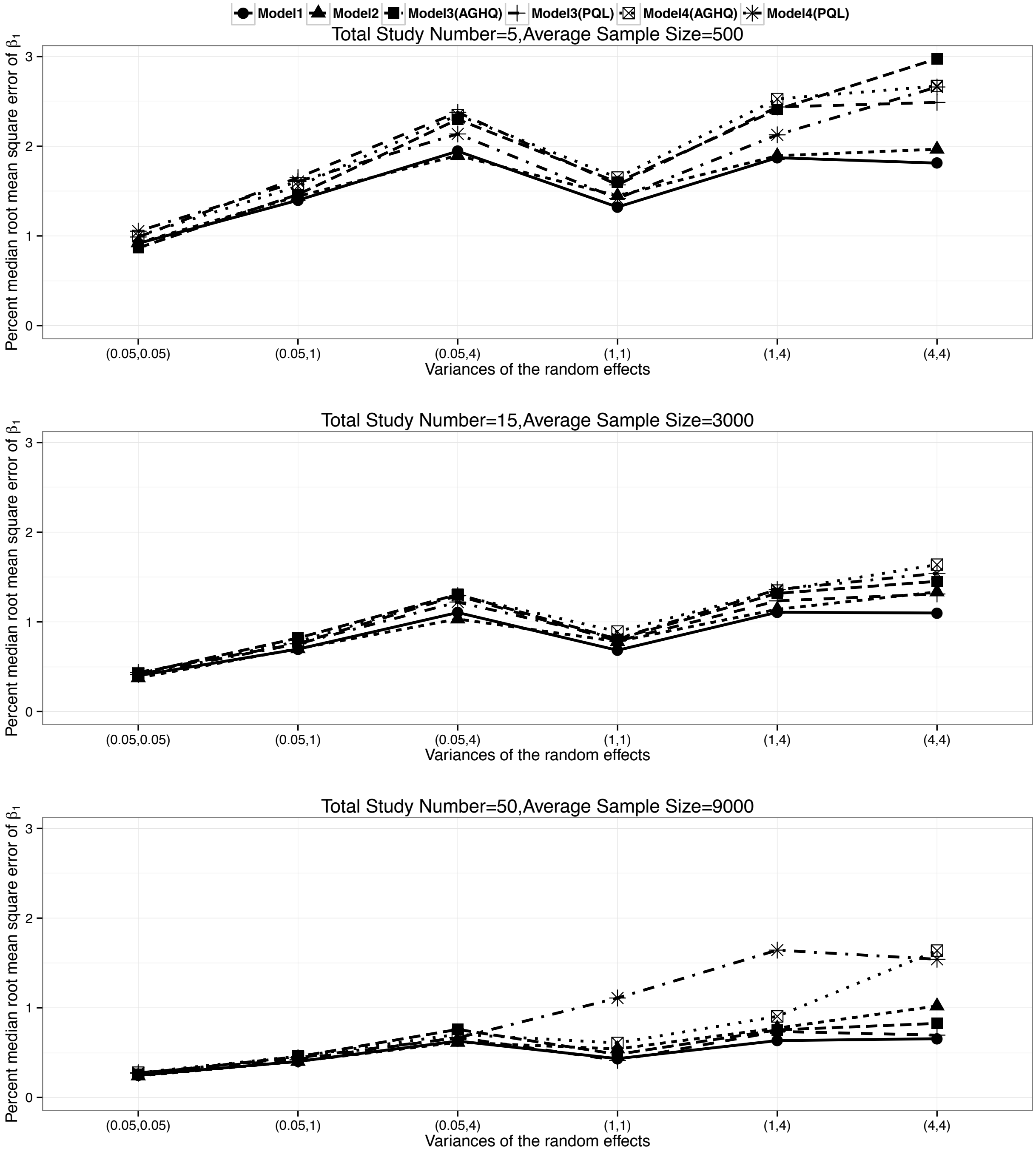

Figure 2 Median percent root mean square error of the pooled treatment effect, β_1, by the number of studies, sample size and methods over increasing random-effect variances: Bivariate two-step (Model1), DeSermonian-Laird two-step (Model2), random intercept one-step (Model3; estimation method: PQL or AGHQ) and stratified intercept one-step (Model4; estimation method: PQL or AGHQ).

Table 4 Percent Coverage[a] (percent convergence rate)[b] for the pooled treatment effect, β_1 by varying number of studies, sample sizes and random-effects variances over increasing random-effect variances: Bivariate two-step (M1), DeSermonian-Laird two-step (M2), random intercept one-step (M3; estimation method: P-PQL or A-AGHQ) and stratified intercept one-step (M4; estimation method: P-PQL or A-AGHQ).

		Random-effects Variances (σ^2,τ^2) [c]					
(K, N)[d]	Model	(0.05, 0.05)	(0.05, 1)	(0.05, 4)	(1,1)	(1,4)	(4,4)
(5,500)	M1	96.6 (100)	87.4 (100)	83.8 (100)	87.5 (100)	86.8 (100)	86.5 (100)
	M2	97.5 (100)	83.9 (100)	81.4 (100)	85 (100)	84 (100)	82 (100)
	M3P	94.9 (99.6)	82.7 (99.7)	80.9 (99.2)	84.6 (99.7)	82.5 (99.3)	83.4 (99.1)
	M3A	79.4 (74)	73.1 (67.7)	71.9 (63.8)	82.4 (48.9)	80.5 (53.8)	89.7 (51.8)
	M4P	99.1 (34.5)	95.6 (41.7)	95.2 (49.4)	96.9 (30.5)	96.5 (39.2)	96.3 (27.9)
	M4A	84.8 (100)	75.4 (99.7)	75 (100)	71.1 (99.9)	72.4 (100)	68.3 (99.9)
(15, 3000)	M1	94.5 (100)	89.5 (100)	92.3 (100)	88 (100)	90.5 (100)	91 (100)
	M2	94.9 (100)	88.5 (100)	91	86.6 (100)	90.3 (100)	86.6 (100)
	M3	91.4 (94.7)	86.9 (84.4)	92.4 (88)	81.6 (95.3)	86.7 (96.3)	86.9 (97.6)
	M3A	92.7 (58)	86.5 (75.6)	87.5 (82.8)	92.6 (84.7)	91.5 (97.3)	92.9 (98.5)
	M4P	96.5 (48.6)	94.9 (56.8)	95.4 (67.5)	97.2 (20.1)	97 (33.8)	99.2 (6.5)
	M4A	86.2 (99.8)	85.8 (100)	89.6 (100)	75.4 (100)	86 (100)	77.6 (100)
(50,9000)	M1	94.6 (100)	93.5 (100)	93.9 (100)	88.7 (100)	90.6 (100)	89.2 (100)
	M2	94.5 (100)	93.3 (100)	93.6 (100)	83.6 (100)	86.1 (100)	77.2 (100)
	M3P	91.4 (94.7)	91.8 (89.1)	95.4 (59.9)	95.7 (23.4)	96.4 (27.3)	93.6 (90.6)
	M3A	97.2 (65.6)	93.4 (90.1)	91.3 (96)	93.4 (100)	93.9 (100)	94.9 (100)
	M4P	99.9 (2.8)	99.8 (6.8)	98.9 (16.3)	99.9 (0.3)	99.9 (0.8)	99.1 (9.5)
	M4A	51.9 (100)	81.8 (100)	91.9 (100)	63.9 (100)	83.3 (100)	89.4 (100)

[a] Percent coverage of β_1 was calculated for each simulated meta-analysis first, and then summarized across meta-analyses. For each combination of data generation parameters, 1000 meta-analyses were generated.

[b] Numerical convergence was only reported for PQL and AGHQ via GLIMMIX procedure in SAS.

[c] σ^2 is the random study-effect variance and τ^2, the random treatment-effect variance

[d] (K,N): (number of studies, total sample size)

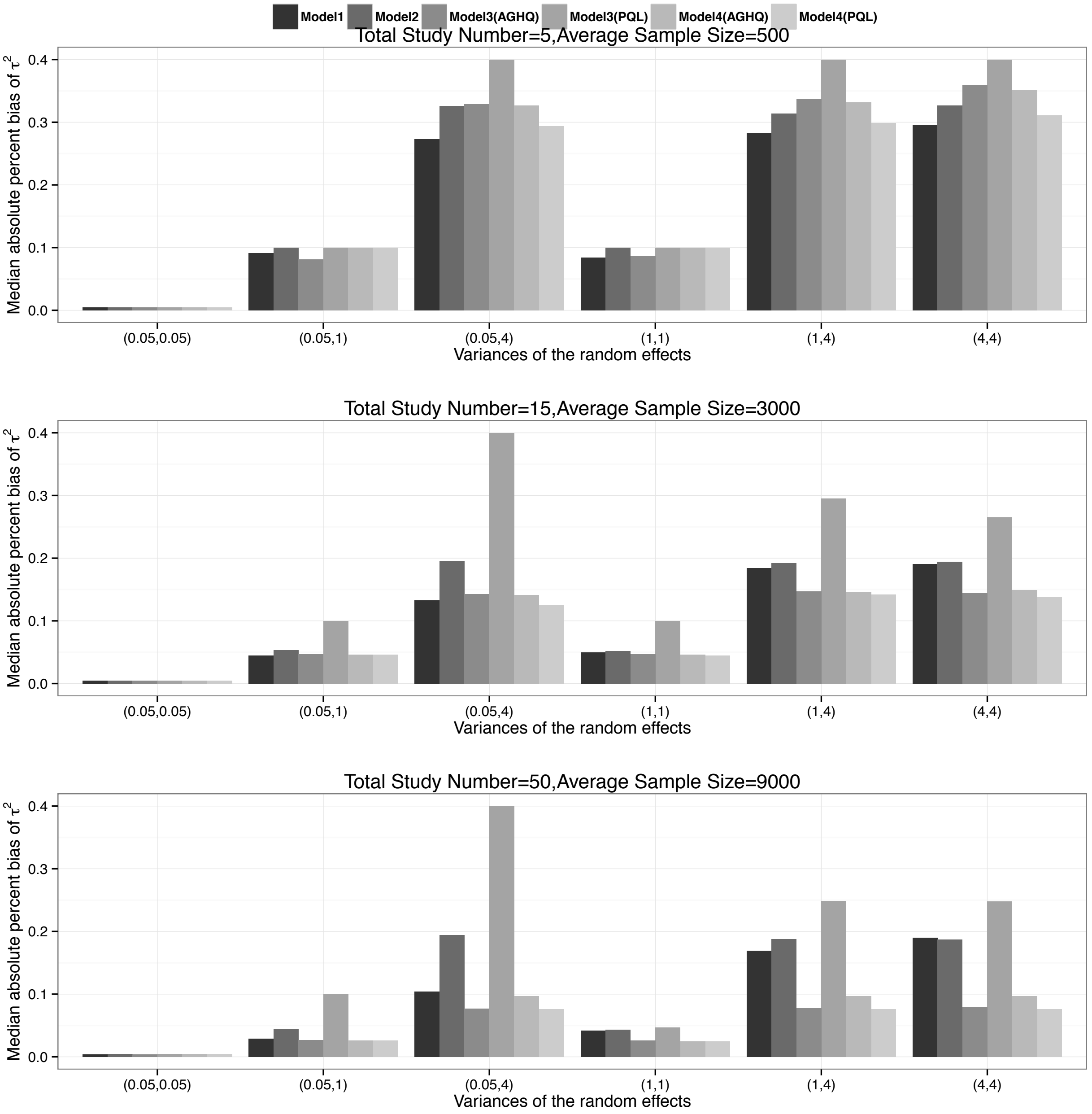

Figure 3 Median absolute bias of the variance component, τ^2 by number of studies, estimation methods and sample sizes over increasing random-effect variances: Bivariate two-step (Model1), DeSermonian-Laird two-step (Model2), random intercept one-step (Model3; estimation method: PQL or AGHQ) and stratified intercept one-step (Model4; estimation method: PQL or AGHQ).

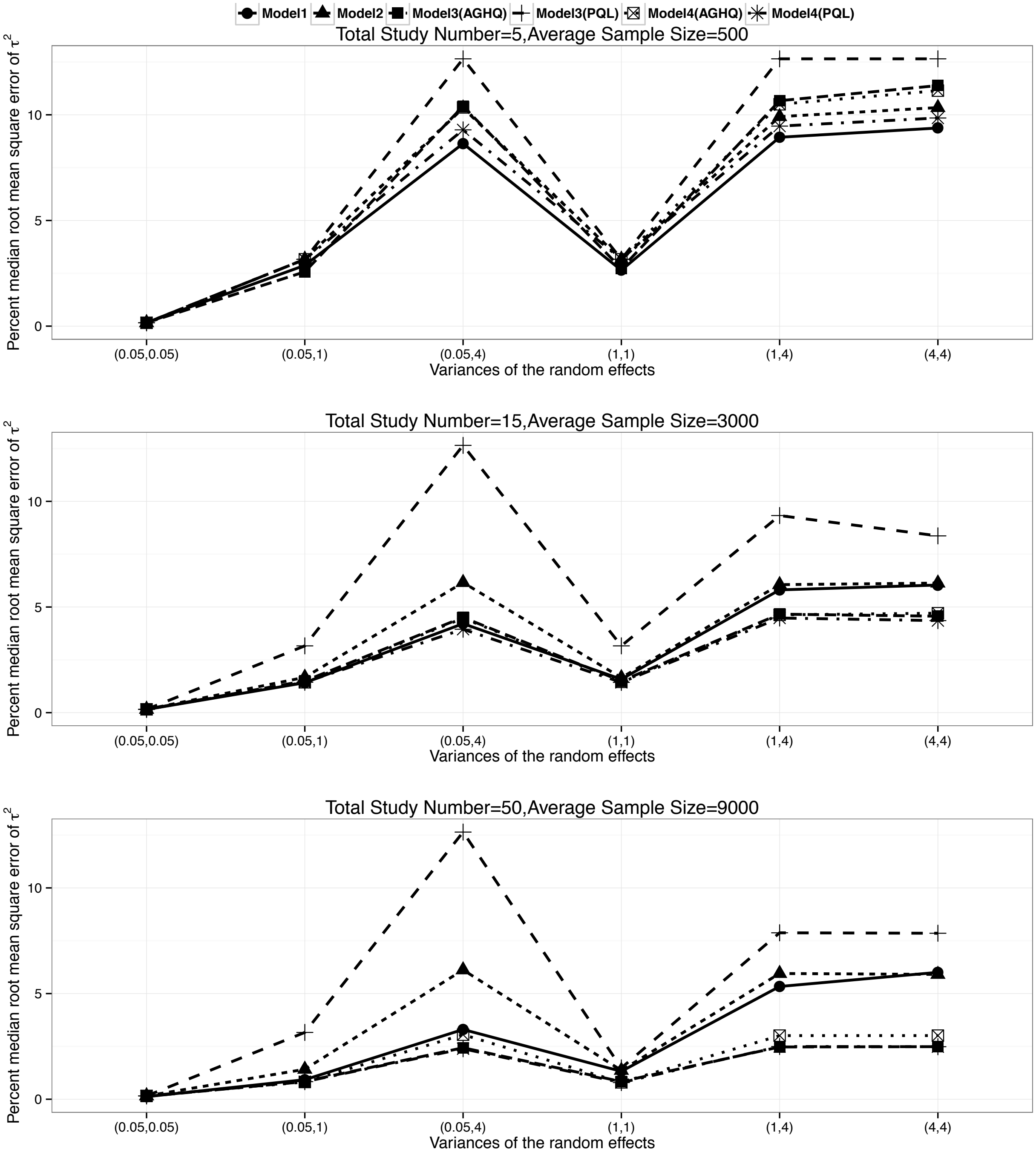

Figure 4 Median percent root mean square error of the variance component, τ^2, by the number of studies, sample size and methods over increasing random-effect variances: Bivariate two-step (Model1), DeSermonian-Laird two-step (Model2), random intercept one-step (Model3; estimation method: PQL or AGHQ) and stratified intercept one-step (Model4; estimation method: PQL or AGHQ).

Table 5 Percent Coverage [15](percent non-zero between study variance estimates) for variance component, τ^2, by the number of studies, sample size and methods over increasing random-effect variances: Bivariate two-step (M1), DeSermonian-Laird two-step (M2), random intercept one-step (M3; estimation method: P-PQL or A-AGHQ) and stratified intercept one-step (M4; estimation method: P-PQL or A-AGHQ).

		Random-effects Variances (σ^2,τ^2) [16]					
(K, N) [17]	Model	(0.05, 0.05)	(0.05, 1)	(0.05, 4)	(1,1)	(1,4)	(4,4)
(5,500)	M1	(52.2)	(68.2)	(89.3)	(82.6)	(93)	(95.3)
	M2	(4.6)	(21.4)	(55.6)	(28.4)	(59.2)	(60.8)
	M3P	100 :8.2 (99.8)	98.1 :5.2 (99.7)	40.9 :1.8 (99.6)	93.8 :4.5 (98.8)	41.6 :3.2 (99.5)	34.9 :7.1 (98.9)
	M3A	27.9 :3.1 (99.9)	58.8 :7.7 (100)	62.8 :12.9 (100)	68.2 :9.3 (99.9)	67.3 :17.5 (99.9)	77.2 :28.5 (100)
	M4P	100 :16.2 (93.6)	99.7 :28.3 (95.8)	93.1 :40.7 (98.9)	96.7 :20.8 (97.4)	89.5 :31.6 (99)	79.4 :20.5 (98.8)
	M4A	52.6 :12.4 (93.1)	74.4 :35 (95.5)	74.8 :53.1 (96.7)	80.6 :35.8 (95.8)	75.4 :51.5 (97.4)	68.9 :46.2 (96.8)
(15, 3000)	M1	(83.4)	(98.9)	(100)	(99.7)	(100)	(100)
	M2	(35.4)	(92)	(98.9)	(91)	(99.2)	(99.1)
	M3P	96.2 :2.6 (99.9)	97.1 :7.4 (99.9)	25.2 :4.2 (100)	94.2 :24.7 (98.2)	0.4 :0.2 (99.2)	0 :0 (99.3)
	M3A	91.3 :22.9 (100)	84.8 :49.4 (100)	72.8 :53.2 (100)	86.5 :69.9 (100)	81.7 :79.2 (100)	83 :81.4 (100)
	M4P	100 :25.3 (94.5)	84.8 :46.4 (99.6)	85.6 :57.6 (100)	91.8 :17.8 (99.6)	87.9 :29.7 (100)	89.2 :5.8 (100)
	M4A	86.1 :36.5 (93.3)	79.6 :73.9 (99.3)	76 :75.8 (100)	82.5 :72.1 (99.9)	75.4:75.1 (100)	76.1:73.9 (99.9)
(50,9000)	M1	(91.6)	(100)	(100)	(100)	(100)	(100)
	M2	(41.6)	(100)	(100)	(100)	(100)	(100)
	M3P	88.9 :0.8 (99.9)	97.1 :6.8 (100)	11.1 :3.4 (99.8)	56.3 :2.7 (98.3)	0 :0 (100)	0 :0 (100)
	M3A	98.7 :54 (100)	89.7 :76.8 (100)	83.1 :78.6 (100)	88.5 :88.4 (100)	91.8 :91.8 (100)	92.1 :92.1 (100)
	M4P	100 :1.2 (97.2)	91.2 :6.2 (100)	87.1 :14.2 (100)	100 :0.3 (100)	100 :0.8 (100)	88 :11.7 (100)
	M4A	99.7 :29.2 (96.2)	87.2 :82.7 (100)	79.3 :79.3 (100)	90 :80.2 (100)	79.8 :79.8 (100)	79.4 :79.4 (100)

[15] Percent coverage of τ^2 was calculated assuming normality for each simulated meta-analysis first, and then summarized across meta-analyses. For each combination of data generation parameters, 1000 meta-analyses were generated. Coverage was reported as a ratio of the subset of cases that excluded meta-analyses where no standard error was estimated, to cases that included these meta-analyses as non-coverage.

[16] σ^2 is the random study-effect variance and τ^2, the random treatment-effect variance

[17] (K, N): (number of studies, total sample size)

Chapter 6 Conclusion

The two manuscripts in this thesis had distinct objectives. In this chapter the results and findings from both are summarized. The first manuscript (Chapter 4) was a systematic review of the statistical methods used in a sample of published individual patient data meta-analyses (IPD-MA) with binary outcomes. In particular, I was interested in the following: (i) whether two-step or one-step methods were frequently used; (ii) how inter-study heterogeneity was calculated and reported; and (ii) if a one-step approach was used, were intercept were permitted to vary across studies as random. In Chapter 5 (Manuscript II), I evaluated via a simulation study the performance of several strategies for analyzing IPD-MA with binary outcomes.

The systematic review (Manuscript 1) included 26 IPD-MA published in 2011 that presented results on binary outcomes. When compared to a previous review [16], I found nearly twice the number of IPD-MA with binary outcome in just one year within this manuscript (14 vs. 26) [56]. Evidently, this review showed that the one-step approach was being used more often in practice than the two-step methods, reflecting it’s flexibility over the two-step, as well as the greater comfort with and availability of software to fit GLMMs. It was also shown that heterogeneity was usually reported (81%) and quantified using the I^2 statistic [29]. However, the statistical approach taken to perform IPD-MA of binary outcomes was often not reported in sufficient detail.

Regarding the simulation results presented in Chapter 5, nearly comparable and unbiased results were obtained for the pooled treatment effect with the one- and two-step methods for scenarios with large study sizes. However, the one-step method outperformed the two-step method in meta-analyses of smaller study size and its method appeared to produce more accurate and precise estimates of the inter-study variance of the treatment effect as compared to the two-step method. It was also reported that correction for attenuation bias and correct model selection can potentially reduce the bias in the one-step method [17 ,20], but the bias in the treatment effect variance estimates for the two-step approach was inherent to the estimation process.

Both the PQL and AGHQ produce mostly unbiased estimates of the pooled treatment effect, particularly when there was low heterogeneity. However, we found that the bias as estimated via the AGHQ procedure was substantial when study sizes were small. In general, coverage was reasonably close to the nominal level with both methods for the pooled treatment effect and as expected, convergence was an issue with the AGHQ procedure for scenarios when the heterogeneity was low.

Also, both PQL and AGHQ produced large biases for the inter-study heterogeneity of the treatment effect, particularly when the variability in the random effects was large. The bias increased as the outcome rate decreased. Overall, AGHQ estimates of the inter-study heterogeneity of the treatment effect were more biased than PQL estimates, with larger RMSE when overall sample was small.

In addition, the simulation results from manuscript II showed the primary advantage of modelling the study-effect as random than stratified when study sizes were small. Hence, the two different sources of heterogeneity can be jointly modelled and estimated. Falsely coercing the study-effect to be common across all studies could lead to inaccurate parameter estimates (pooled treatment effect and its inter-study heterogeneity).

Finally, the number of studies (result not shown) had the greatest impact on the performance indicators than the other generation parameters. Meta-analyses of small studies were observed to pose severe challenges, however, in these cases the PQL procedure tended to perform better than the AGHQ method. We recommend that the one-step method be used in IPD-MA of binary outcome when study sizes are small, as it was most appropriate for these scenarios. The two-step method does not produce biased pooled treatment effects, but frequently underestimates the inter-study heterogeneity.

Future work in the area of IPD-MA of binary outcome is necessary, particularly; the relative merits of the Bayesian approach (that allows non-normal distribution to also be specified for the random effects) offers some substantial advantages over the conventional likelihood approaches [68 ,69]. In addition, unresolved issues concerning methods performance for the pooling of observational studies and addressing the known biases that frequently occur in these designs (clinical differences between study population, misclassification of subjects, controlling for confounding variables etc.) [70 ,71].

LIST OF REFERENCES

1. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials 1986;**7**(3):177-88
2. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ (Clinical research ed) 2010;**340**:c221 doi: 10.1136/bmj.c221[published Online First: Epub Date]|.
3. Riley RD. Commentary: like it and lump it? Meta-analysis using individual participant data. International journal of epidemiology 2010;**39**(5):1359-61 doi: 10.1093/ije/dyq129[published Online First: Epub Date]|.
4. Blettner M, Sauerbrei W, Schlehofer B, et al. Traditional reviews, meta-analyses and pooled analyses in epidemiology. International journal of epidemiology 1999;**28**(1):1-9 doi: 10.1093/ije/28.1.1[published Online First: Epub Date]|.
5. Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? Lancet 1993;**341**(8842):418-22
6. Sud S, Douketis J. The devil is in the details...or not? A primer on individual patient data meta-analysis. Evidence-based medicine 2009;**14**(4):100-1 doi: 10.1136/ebm.14.4.100[published Online First: Epub Date]|.
7. Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. Journal of clinical epidemiology 2007;**60**(5):431-9 doi: 10.1016/j.jclinepi.2006.09.009[published Online First: Epub Date]|.
8. Lambert PC, Sutton AJ, Abrams KR, et al. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. Journal of clinical epidemiology 2002;**55**(1):86-94
9. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. Evaluation & the health professions 2002;**25**(1):76-97
10. Simmonds MC, Higgins JP. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. Statistics in medicine 2007;**26**(15):2982-99 doi: 10.1002/sim.2768[published Online First: Epub Date]|.
11. Higgins JP, Whitehead A, Turner RM, et al. Meta-analysis of continuous outcome data from individual patients. Statistics in medicine 2001;**20**(15):2219-41 doi: 10.1002/sim.918[published Online First: Epub Date]|.
12. Berlin JA, Santanna J, Schmid CH, et al. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Statistics in medicine 2002;**21**(3):371-87
13. Stewart LA, Clarke MJ. Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. Statistics in medicine 1995;**14**(19):2057-79
14. Mathew T, Nordstrom K. On the Equivalence of Meta-Analysis Using Literature and Using Individual Patient Data. BIOM Biometrics 1999;**55**(4):1221-23

15. Riley RD, Lambert PC, Staessen JA, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Statistics in medicine 2008;**27**(11):1870-93 doi: 10.1002/sim.3165[published Online First: Epub Date]|.
16. Simmonds MC, Higgins JP, Stewart LA, et al. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clinical trials (London, England) 2005;**2**(3):209-17
17. Debray TPA, Moons KGM, Abo-Zaid GMA, et al. Individual Participant Data Meta-Analysis for a Binary Outcome: One-Stage or Two-Stage? PLoS ONE 2013;**8**(4):e60650 doi: 10.1371/journal.pone.0060650[published Online First: Epub Date]|.
18. Schmid CH, Stark PC, Berlin JA, et al. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. Journal of clinical epidemiology 2004;**57**(7):683-97 doi: 10.1016/j.jclinepi.2003.12.001[published Online First: Epub Date]|.
19. Mathew T, Nordstrom K. Comparison of one-step and two-step meta-analysis models using individual patient data. Biometrical journal Biometrische Zeitschrift 2010;**52**(2):271-87 doi: 10.1002/bimj.200900143[published Online First: Epub Date]|.
20. Stewart GB, Altman DG, Askie LM, et al. Statistical Analysis of Individual Participant Data Meta-Analyses: A Comparison of Methods and Recommendations for Practice. PLoS ONE 2012;**7**(10):e46042 doi: 10.1371/journal.pone.0046042[published Online First: Epub Date]|.
21. Borenstein M, Hedges LV, Higgins JP, et al. *Introduction to Meta-Analysis*. 1 ed. Chichester, UK.: John Wiley & Sons, Ltd., 2009.
22. Stram DO. Meta-Analysis of Published Data Using a Linear Mixed-Effects Model. Biometrics 1996;**52**(2):536-44 doi: 10.2307/2532893[published Online First: Epub Date]|.
23. Peto R. Why do we need systematic overviews of randomized trials? Statistics in medicine 1987;**6**(3):233-44
24. Begg CB, Pilote L. A model for incorporating historical controls into a meta-analysis. Biometrics 1991;**47**(3):899-906
25. Torri V, Simon R, Russek-Cohen E, et al. Statistical model to determine the relationship of response and survival in patients with advanced ovarian cancer treated with chemotherapy. Journal of the National Cancer Institute 1992;**84**(6):407-14
26. Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews. Statistics in medicine 2002;**21**(11):1503-11 doi: 10.1002/sim.1183[published Online First: Epub Date]|.
27. Bowden J, Tierney JF, Copas AJ, et al. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC Med Res Methodol 2011;**11**:41 doi: 10.1186/1471-2288-11-41[published Online First: Epub Date]|.
28. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. Statistics in medicine 1988;**7**(8):889-94 doi: 10.1002/sim.4780070807[published Online First: Epub Date]|.
29. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in medicine 2002;**21**(11):1539-58 doi: 10.1002/sim.1186[published Online First: Epub Date]|.

30. Rucker G, Schwarzer G, Carpenter JR, et al. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol 2008;**8**:79 doi: 10.1186/1471-2288-8-79[published Online First: Epub Date]|.
31. McCulloch CE, Searle SR. *Generalized, Linear, and Mixed Models*. New York: Wiley, 2000.
32. Molenberghs; G, Verbeke G. *Models for Discrete Longitudinal Data*. New York: Springer, 2005.
33. Diggle; P, Heagerty; P, Liang; K-Y, et al. *Analysis of Longitudinal Data*. Second ed. New York: Oxford University Press, 2002.
34. Turner RM, Omar RZ, Yang M, et al. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Statistics in medicine 2000;**19**(24):3417-32
35. Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association 1993;**88**(421):9-25 doi: 10.2307/2290687[published Online First: Epub Date]|.
36. Pinheiro JC, Bates DM. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. Journal of Computational and Graphical Statistics 1995;**4**(1):12-35 doi: 10.2307/1390625[published Online First: Epub Date]|.
37. Best NG, Spiegelhalter DJ, Thomas A, et al. Bayesian Analysis of Realistically Complex Models. Journal of the Royal Statistical Society Series A (Statistics in Society) 1996;**159**(2):323-42 doi: 10.2307/2983178[published Online First: Epub Date]|.
38. Jang W, Lim J. A Numerical Study of PQL Estimation Biases in Generalized Linear Mixed Models Under Heterogeneity of Random Effects. Communications in Statistics - Simulation and Computation 2009;**38**(4):692-702 doi: 10.1080/03610910802627055[published Online First: Epub Date]|.
39. Wolfinger R, O'Connell M. Generalized linear mixed models a pseudo-likelihood approach. Journal of Statistical Computation and Simulation 1993;**48**(3-4):233-43 doi: 10.1080/00949659308811554[published Online First: Epub Date]|.
40. Goldstein H. Nonlinear multilevel models, with an application to discrete response data. Biometrika 1991;**78**(1):45-51 doi: 10.1093/biomet/78.1.45[published Online First: Epub Date]|.
41. Breslow NE, Lin X. Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion. Biometrika 1995;**82**(1):81-91 doi: 10.2307/2337629[published Online First: Epub Date]|.
42. Lin X, Breslow NE. Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion. Journal of the American Statistical Association 1996;**91**(435):1007-16 doi: 10.2307/2291720[published Online First: Epub Date]|.
43. Shun Z, McCullagh P. Laplace Approximation of High Dimensional Integrals. Journal of the Royal Statistical Society Series B (Methodological) 1995;**57**(4):749-60 doi: 10.2307/2345941[published Online First: Epub Date]|.
44. Shun Z. Another Look at the Salamander Mating Data: A Modified Laplace Approximation Approach. Journal of the American Statistical Association 1997;**92**(437):341-49 doi: 10.2307/2291479[published Online First: Epub Date]|.
45. Raudenbush SW, Yang M-L, Yosef M. Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. Journal of Computational and Graphical Statistics 2000;**9**(1):141-57 doi: 10.2307/1390617[published Online First: Epub Date]|.
46. Demidenko E. *Mixed Models: Theory and Applications*. New York: Wiley, 2004.

47. Hedeker D, Gibbons RD. *Longitudinal Data Analysis*. New York: Wiley, 2006.
48. Lee Y, Nelder JA, Pawitan Y. *Generalized Linear Models with Random Effect: Unified Analysis with H-likelihood*. London: Chapman and Hall, 2006.
49. Abramowitz; M, Stegan I. *Handbook of Mathematical Functions*. Washington, DC: National Bureau of Standards, 1964.
50. Naylor JC, Smith AFM. Applications of a Method for the Efficient Computation of Posterior Distributions. Journal of the Royal Statistical Society Series C (Applied Statistics) 1982;**31**(3):214-25 doi: 10.2307/2347995[published Online First: Epub Date]|.
51. Liu Q, Pierce DA. A Note on Gauss-Hermite Quadrature. Biometrika 1994;**81**(3):624-29 doi: 10.2307/2337136[published Online First: Epub Date]|.
52. Diaz RE. Comparison of PQL and Laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomised trials. Computational Statistics & Data Analysis 2007;**51**(6):2871-88 doi: http://dx.doi.org/10.1016/j.csda.2006.10.005%5Bpublished Online First: Epub Date]|.
53. Callens M, Croux C. Performance of likelihood-based estimation methods for multilevel binary regression models. Journal of Statistical Computation and Simulation 2005;**75**(12):1003-17 doi: 10.1080/00949650412331321070[published Online First: Epub Date]|.
54. Capanu M, Gönen M, Begg CB. An assessment of estimation methods for generalized linear mixed models with binary outcomes. Statistics in medicine 2013;**32**(26):4550-66 doi: 10.1002/sim.5866[published Online First: Epub Date]|.
55. Rondeau V, Michiels S, Liquet B, et al. Investigating trial and treatment heterogeneity in an individual patient data meta-analysis of survival data by means of the penalized maximum likelihood approach. Statistics in medicine 2008;**27**(11):1894-910 doi: 10.1002/sim.3161[published Online First: Epub Date]|.
56. Thomas D, Radji S, Benedetti A. Systematic review of methods for individual patient data meta- analysis with binary outcomes. BMC Med Res Methodol 2013
57. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in medicine 2002;**21**(4):589-624 doi: 10.1002/sim.1040[published Online First: Epub Date]|.
58. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2009;**172**(4):789-811 doi: 10.1111/j.1467-985X.2008.00593.x[published Online First: Epub Date]|.
59. Chen H, Manning AK, Dupuis J. A Method of Moments Estimator for Random Effect Multivariate Meta-Analysis. Biometrics 2012;**68**(4):1278-84 doi: 10.1111/j.1541-0420.2012.01761.x[published Online First: Epub Date]|.
60. Hardy RJ, Thompson SG. A LIKELIHOOD APPROACH TO META-ANALYSIS WITH RANDOM EFFECTS. Statistics in medicine 1996;**15**(6):619-29 doi: 10.1002/(SICI)1097-0258(19960330)15:6<619::AID-SIM188>3.0.CO;2-A[published Online First: Epub Date]|.
61. Littell RC, Milliken GA, Stroup WW, et al. Cary, NC: SAS System for Mixed Models, 1996:633.
62. SAS/STAT(R) 9.2 User's Guide, Second Edition. Secondary SAS/STAT(R) 9.2 User's Guide, Second Edition 2015. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm - statug_glimmix_a0000001460.htm.

63. Abo-Zaid G, Guo B, Deeks JJ, et al. Individual participant data meta-analyses should not ignore clustering. Journal of clinical epidemiology 2013;**66**(8):865-73.e4 doi: http://dx.doi.org/10.1016/j.jclinepi.2012.12.017%5Bpublished Online First: Epub Date]|.

64. Mathew T, Nordström K. Comparison of One-Step and Two-Step Meta-Analysis Models Using Individual Patient Data. Biometrical Journal 2010;**52**(2):271-87 doi: 10.1002/bimj.200900143[published Online First: Epub Date]|.

65. Moineddin R, Matheson FI, Glazier RH. A simulation study of sample size for multilevel logistic regression models. BMC Med Res Methodol 2007;**7**:34 doi: 10.1186/1471-2288-7-34[published Online First: Epub Date]|.

66. Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge; New York: Cambridge University Press, 2007.

67. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in medicine 2004;**23**(20):3105-24 doi: 10.1002/sim.1875[published Online First: Epub Date]|.

68. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. Statistics in medicine 2001;**20**(3):453-72 doi: 10.1002/1097-0258(20010215)20:3<453::AID-SIM803>3.0.CO;2-L[published Online First: Epub Date]|.

69. Browne; WJ, Draper D. A comparison of Bayesian and likelihood- based methods for fitting multilevel models. . Bayesian Analysis 2006;**1**:473-514

70. Sutton AJ. *Methods for meta-analysis in medical research*. Chichester; New York: J. Wiley, 2000.

71. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. Journal of clinical epidemiology 1991;**44**(2):127-39