MAKING THE MOST OF A FEW

SMALL POPULATION CLINICAL TRIALS

Konstantinos Pateras

Making the most of a few small population clinical trials

Konstantinos Pateras

Making the most of a few small population clinical trials

PhD thesis.
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands.

ISBN:	978-90-393-7328-6
Author:	Konstantinos Pateras
Cover Photo:	Vasilis Pateras, Instagram[@vasilis_pateras]
Cover Design:	Giorgos Markou, [www.behance.net/GeorgeMArkou]
Printing:	Matura Edition, Korydallos, Greece, [maturapress@gmail.com]

The Julius Center for Health Sciences and Primary Care financially supported the publication of this thesis.

Making the most of a few small population clinical trials

Αξιοποιώντας στο έπακρο λίγες κλινικές δοκιμές σε μικρούς πληθυσμούς

(με περίληψη στα Ελληνικά)

Optimaal profiteren van enkele klinische onderzoeken met kleine populaties

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 28 oktober 2020 des ochtends te 11.00 uur

door

Konstantinos Pateras

geboren op 28 maart 1985
te Athene, Griekenland

Promotor: Prof. dr. C.B. Roes
Copromotor: Dr. S. Nikolakopoulos

Dit proefschrift werd (mede) mogelijk gemaakt met financiële steun van de European Union's seventh framework programme (FP7-HEALTH-2013-INNOVATION-1, Grant-Agreement No. 603160).

In girum imus nocte et consumimur igni

Manuscripts included in this thesis

Chapter 2a	I van der Tweel, **K Pateras**, GCM van Baal, KCB Roes. A review of frequentist methods for combining results of series of trials. [*Internal Asterix Report*]
Chapter 2b	**K Pateras**, L Spineli, KCB Roes. Bayesian evidence synthesis for combining results of series of a few small trials. [*Thesis exclusive*]
Chapter 3	**K Pateras**, S Nikolakopoulos, KCB Roes. Data generating models of dichotomous outcomes: Heterogeneity in simulation studies for a random-effects meta-analysis. [*Statistics in Medicine. 2018 ; 37(7): 1115-1124*]
Chapter 4	**K Pateras**, S Nikolakopoulos, D Mavridis, KCB Roes. Interval estimation of the overall treatment effect in a meta-analysis of a few small studies with zero events. [*Contemporary Clinical Trials Communications. 2018 ; (9): 98–107*]
Chapter 5	**K Pateras**, S Nikolakopoulos, KCB Roes. Prior distributions for variance parameters in a sparse-event meta-analysis of a few small trials. [*Pharmaceutical Statistics. 2020*]
Chapter 6	**K Pateras**, S Nikolakopoulos, KCB Roes. Borrowing strength from early phase outcomes in orphan drug development. [*Under revision*]
Chapter 7	L Spineli, C Kalyvas and **K Pateras**. Participants' outcomes gone missing within a network of interventions: Bayesian modeling strategies. [*Statistics in Medicine. 2019 ; 38: 3861–3879*]

Contents

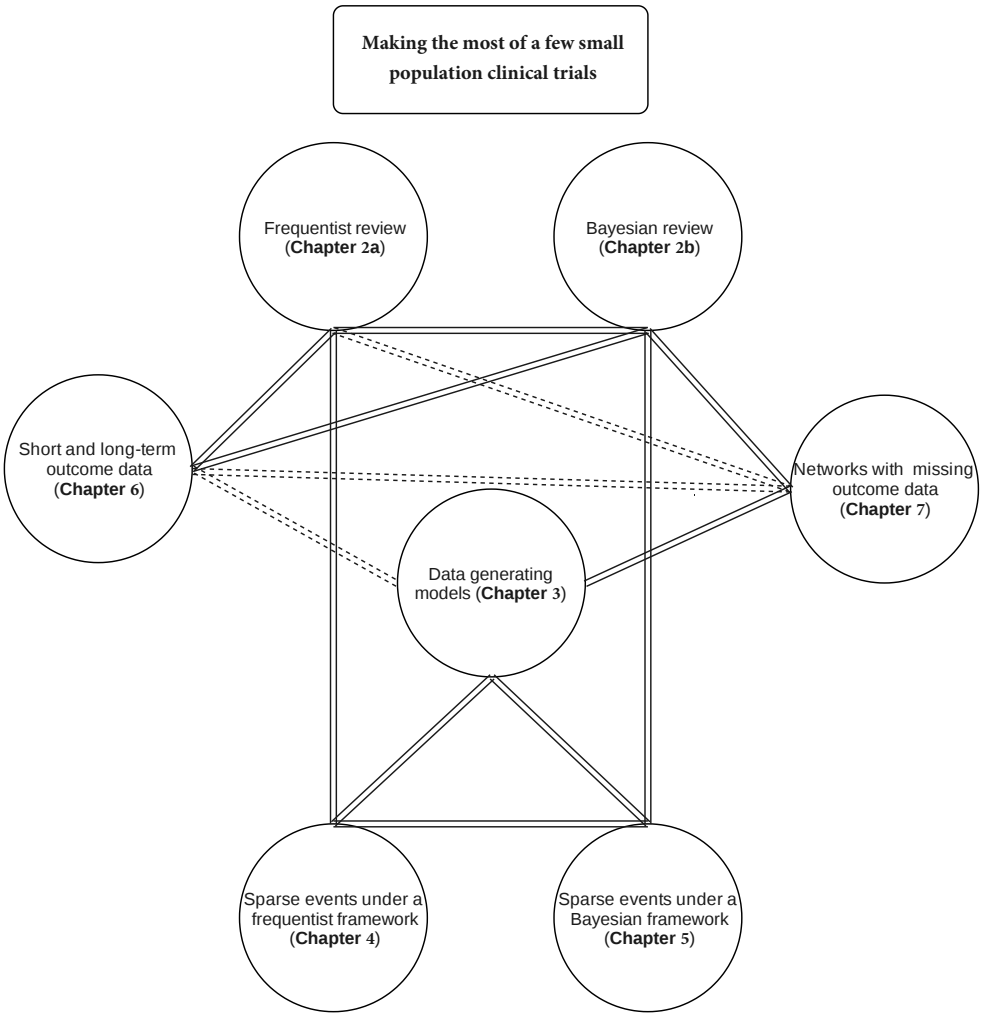

Figure 1: Chart of conceptual connection of chapters. Solid lines refer to direct connections, while dashed lines refer to indirect connections of chapters.

Chapter 1

General introduction

1.1 Evidence derived from small populations

Clinical research in (very) rare diseases is challenging. One of the main reasons is the small populations of patients suffering from each individual rare disease. However, since there are more that 7000 known rare diseases, a 7% of people in total will suffer from a rare disease at least once during their lifetime [1]. Small patient populations could also emerge from stratification of "common" diseases" on the basis of some biomarker, e.g. genomic markers. Interventional research in small populations faces a substantial challenge; the limited number of patients to be included in clinical trials, which naturally impacts the level of generated evidence.

Randomized controlled clinical trials are considered the gold standard for assessing the efficacy and safety of (innovative) medical interventions. Both exploratory and confirmatory clinical trials in small populations risk producing non-conclusive evidence for the efficacy of a treatment. Nonetheless, randomized controlled clinical trials are the preferred means for the evaluation of treatments in rare diseases. This paradigm is in place both to ensure similar high standards of evidence as common disease treatments and to protect the quality of approved interventions [2].

The recognition of the shortcomings of available statistical methods to evaluate rare disease interventions initiated efforts that provided approaches tailored specifically for small populations [3, 4]. Nonetheless, additional suitable and innovative approaches need to be developed for evaluating observational and randomized evidence for rare diseases and this ongoing attempt is further recognised by both the research community and regulatory authorities [2, 5]. Following this need, the European Commission supported three international research projects towards this direction; namely "Advances in Small Trials Design for Regulatory Innovation and Excellence" [6], "Integrated Design and Analysis of Small Population Group Trials" [7] and "Innovative Methodology for Small Population Research" [8].

1.2 Evidence synthesis in small populations

In small populations, the very low prevalence of each disease often generates a limited pool of small and possibly underpowered clinical trials. In such cases a synthesis of data across

clinical trials can generate a more informed, generalizable and usually more powerful result. Utilizing all relevant available exploratory and/or confirmatory clinical trials, such a synthesis can be conducted via a meta-analysis [9] or a network meta- analysis [10]. The synthesized trials can differ in terms of methodological and clinical characteristics such as the study designs, the trial populations or the definition of outcomes for each trial. These differences may lead to variability in the true underlying effect which will manifest itself in between-study variance – heterogeneity – of the observed effects. Quantifying heterogeneity in a meta-analysis of a few small clinical trials poses a challenge for most statistical approaches. In fact, the synthesis of two heterogeneous studies was recently presented as an unresolved issue within a frequentist framework [11] and initiated the exploration of Bayesian evidence synthesis approaches as more robust alternatives [12, 13].

Bayesian inference is often debated due to its subjective nature. Nevertheless, it offers an appealing alternative for the synthesis of few and small clinical trials. The consideration of Bayesian methods in small populations is recommended by international guidelines [2, 14, 15] and published research [12, 13]. Nonetheless, Bayesian meta-analysis of a few small trials risks resulting in prior-driven inferences and unknown frequentist operational characteristics. This calls for rigorous evaluation of these characteristics during the design and meta-analysis of a series of trials.

Either through a frequentist or a Bayesian framework, synthesis of all relevant data could offer further insights. Under a frequentist framework, standard operational statistical characteristics - such as: (1) type I error control at 5%, (2) 95% coverage of confidence intervals and (3) minimal bias - are difficult to achieve without becoming (very) conservative due to the small number of available trials. Likewise, under a Bayesian framework, inferences can quickly become prior-driven. In such a small population setting, the sparsity of available information makes existing frequentist and Bayesian evidence synthesis approaches problematic.

In this thesis, I recognize and discuss the following themes: (1) small population clinical trials may lead to deviations from asymptotic assumptions, (2) small population clinical trials may lead to excessive zero events, (3) small population clinical trials may hamper estimation of heterogeneity, (4) in small populations, evidence generation may benefit from the combination

of exploratory and confirmatory clinical trials and (5) informative missing outcome data may impact meta-analysis, especially in the case of network meta-analysis, where study designs are even more diverse.

1.3 Special issues in evidence synthesis of small populations

The methodological reviews of Chapter 2a and 2b identify a possible lack of available sophisticated approaches and provide possible directions for further research on statistical evidence synthesis. These directions partially shaped the contents of this thesis.

Small population clinical trials may lead to deviations from assumptions that are based on asymptotic approximations, due to the limited sample size. A commonly encountered assumption is that of the normality of test statistics, an assumption which most current synthesis approaches are built upon. This assumption becomes more problematic in random-effects models that assumes two levels of normality, one between trials and one between patients. For dichotomous outcomes, most methodological research studies on meta-analysis in rare diseases discuss and compare statistical approaches under a normal approximation of the binomial distribution [11, 13, 16, 17]. Such an assumption is not suitable and often brakes down in the case of small samples and/or small number of events. When these simulation studies deviate from the normal approximation for the binomial distribution, even the set-up of the data generation mechanism of their simulation becomes non trivial and each study often deploys a different data generating model. Such practices may result in recommendations that are based on different assumptions across several investigations. Chapter 3 focuses on the consequences of applying different data generating models for the evaluation of a random-effects meta-analysis for small populations with binomial outcome data.

Small population clinical trials may lead to zero events in one or more of the treatment arms if the outcome is a clinical event, which induce additional methodological challenges for a meta-analysis of these trials. Binary endpoints usually relate to events of clinical importance and are commonly used in clinical trials. Due to the limited trial sample sizes in this setting, zero cells are more likely to be observed in at least one of the treatment arms. As the number of zero cells increases, the unbiased estimation of heterogeneity becomes infeasible and leads to improper (interval) estimation of the overall treatment effect. Such methodological challenges

can occur both under a frequentist (Chapter 4) and a Bayesian (Chapter 5) meta-analysis of a few small trials with sparse events.

Small population clinical trials may lead to the informal synthesis of exploratory and confirmatory clinical trials, that do not both include the same clinical outcomes. In Chapter 6 I discuss the sample-based selection bias which appears when short-term outcomes from a short-term exploratory trial are utilized as supportive evidence to the primary long-term outcome of the current long-term confirmatory trial. I provide solutions for synthesis of such trials that reduce or eliminate this selection bias.

Small population clinical trials may lead to missing outcomes that can be either missing-at-random or missing-not-at-random. Missingness in a meta-analysis is often conveniently modelled through the missing-at-random assumption. In Chapter 7 various modelling options are suggested and explored for informative binary missing outcome data in a Bayesian meta-analytical network of interventions under the missing-at-random assumption. Moreover, building on the work of Chapter 3, a novel straightforward generalization of a data generating mechanism for an network meta-analysis with informative missing outcome data is developed.

The thesis concludes with a general discussion in Chapter 8.

Chapter 2a

A review of frequentist methods for combining results of series of trials

I van der Tweel

K Pateras

GCM van Baal

KCB Roes

Asterix internal report

Abstract

A randomized controlled trial is considered the gold standard in clinical research, also in rare diseases. However, in small populations, single large scale well-powered trials are often not possible. For valid decision-making, both on efficacy and on safety, evidence generated from series of trials could be exploited. We conducted a review over a five-year period to identify relevant methodology on combining results of series of trials. Out of a total of 8183 papers found, 61 papers were included and summarized in this review. Its focus is on frequentist methodology. Most papers deal with meta-analyses on aggregated data. Only few papers discuss multivariate outcomes. We categorized the relevant methods according to the type of (meta-) analysis. Only a few papers dealt directly with series of trials in small populations. The results of the review lead to some directions for further investigation on evidence-based decision-making from a (small) series of trials in small populations.

Introduction

Recently, the European Union funded the ASTeRIx project: Advances in Small Trials dEsign for Regulatory Innovation and eXcellence (see also O'Connor and Hemmings [18]); to develop and implement innovative statistical methodologies for the evaluation of orphan drug treatments with clinical trials. Please note that the list of references is split into two parts: 1) references to other publications and 2) references to papers in the review ([‡]).

Rare diseases influence only a small part of the human population with a prevalence below 5 per 10,000 people in the European Community [1]. To evaluate the effect of a (new) intervention, a randomized controlled trial (RCT) is considered the gold standard also in rare diseases. However, large-scale well-powered RCTs are often not possible. To obtain sufficient, valid evidence for decision-making, both on efficacy and on safety, alternative methodological approaches have to be sought. Existing guidance [5, 19, 20] discusses and recommends designs that are suitable for single trials in small populations. To cope with the problem of small numbers of patients available for a single trial, evidence generated from series of trials in small populations could be exploited.

We performed a review to identify new frequentist methods for series of trials. We also focused on existing methods for large-scale diseases that might be applicable in small populations. In the Results section, we categorize the relevant methods according to the type of (meta-)analysis. In the Discussion section, we will assess the usefulness and limitations of the described methods in a small series of small clinical trials.

Methods

We conducted a review to identify relevant methodology on combining results of series of trials as published between 1-1-2009 and 31-12-2013. Eligible studies were identified with several search strategies. First, we created a list of landmark papers, i.e. specific papers we wanted to be found and included in our final set of papers. Then we created a search strategy for PubMed. Because of the methodological nature of our review, clearly papers were missed by searching PubMed alone. We then extended our search to Web of Science (Science Citation Index (SCI)), Scopus, JSTOR and, lastly, the Cochrane Library (see Appendix for the search strategies). The search resulted in 8183 papers, of which 1031 from PubMed, 2438 from

Scopus, 2230 from Web of Science, 2436 from JSTOR and 48 from Cochrane (considering only methodological articles) (see Figure 2a.1 for the flow diagram of the search strategies). First of all, duplicates were removed. Then papers were excluded by journal, title and abstract based on their methodological relevance for the review. Two reviewers (KP, IvdT) independently scored the 358 articles from the remaining studies by judging title and abstract to include (I), probably include (PI), probably exclude (PE) and to exclude (E).

Articles with a concordant score from both reviewers were either included (I or PI) or excluded (PE or E) from the pool. The discordant 52 ones were discussed with a third independent colleague (GCMvB). From these 52, 38 were excluded and 14 were included into the final set of articles. Some papers with no abstract were considered for full reading; when they consisted of letters to the editor or commentaries on papers not included in the review they were excluded. One review article to which we could not get access was excluded. We only included papers written in English. Excluded were on-line abstracts only, books or book chapters, papers describing applications of meta-analyses and papers on n-of-1 trials.

Finally, we split the included papers into those on Bayesian methods and those on frequentist methods. The frequentist methods will be described and summarized in this review; the Bayesian methods, including most of the papers on network meta-analysis are discussed in Chapter 2b. Papers comparing frequentist and Bayesian methods were discussed in both reviews.

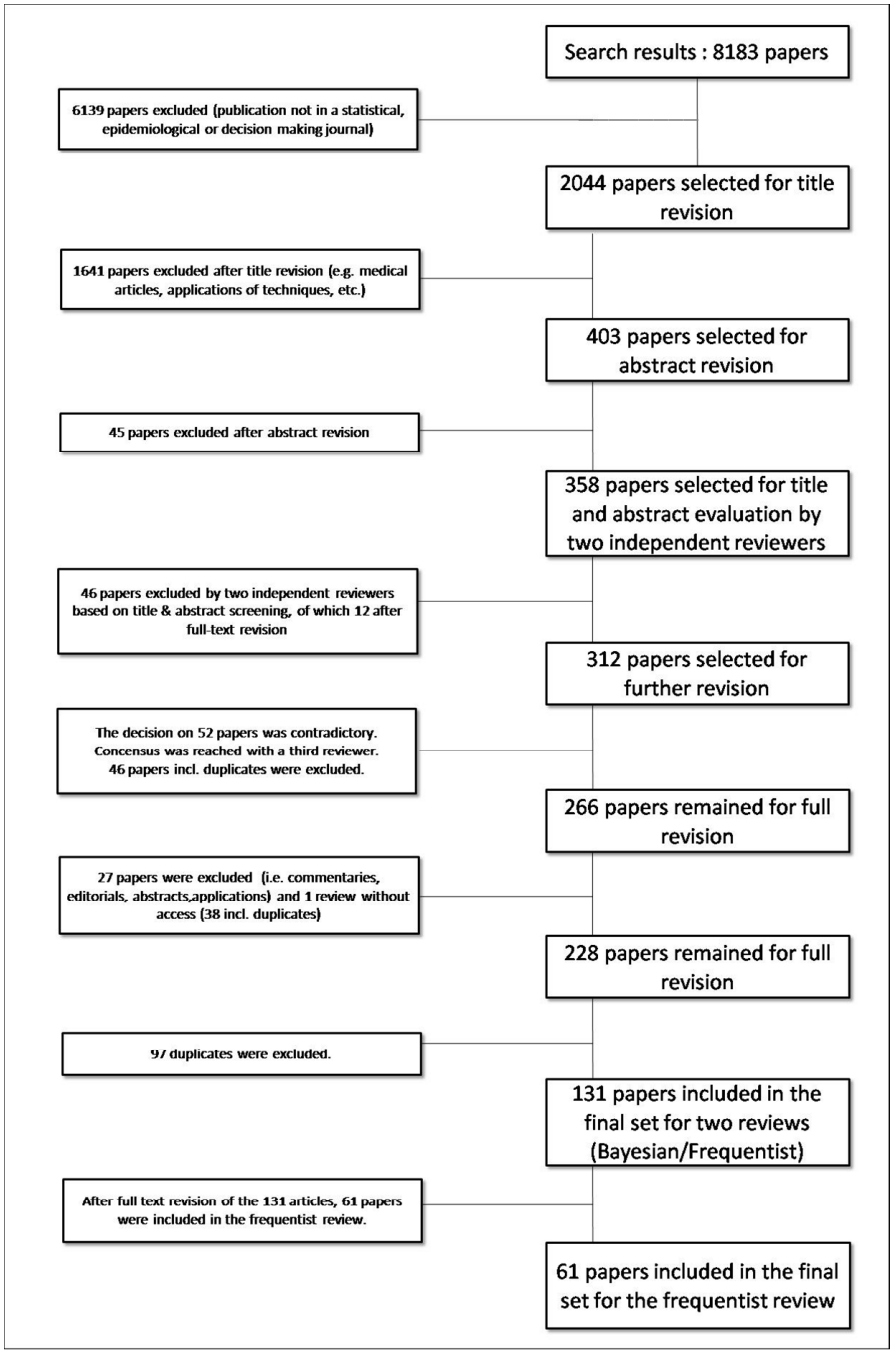

Figure 2a.1: Chapter 2a - Flow diagram of the search strategy

Results

Out of an initial 8183 papers, 61 were included in this review. Some descriptive characteristics of the reviewed papers are presented in Table 2a.1. Most papers deal with methods on summarized or aggregated data. Only few papers discuss multivariate outcomes. A number of papers describe methods for more than one outcome type or discuss various forms of meta-analysis.

Table 2a.1: Characteristics of the articles included in the review. *MA = meta-analysis, CMA = cumulative meta-analysis, TSA = trial sequential meta-analysis, SMA = sequential meta-analysis, PMA = prospective meta-analysis, NMA = network meta-analysis, MTC = mixed treatment comparison, AD = aggregated data, IPD = individual patient data, n.s. = not specified, n.a. = not applicable.

Characteristic	Category*	n
Year of publication	2009	15
	2010	10
	2011	15
	2012	8
	2013	13
Type of MA	General MA	37
	CMA	14
	TSA	13
	SMA	6
	PMA	9
	NMA/MTC	8
	Case series	3
Input data	AD	49
	IPD	7
	AD+IPD	5
Outcome	Univariate	50
	Multivariate	8
	Uni- and Multivariate	3
Type of outcome	Dichotomous	41
	Continuous	14
	Time-to-event	16
	Other	2
	n.s.	10
Approach	Frequentist	46
	Frequentist and Bayes	9
	Hybrid	5
	Empirical Bayes	1
Program	WinBugs	11
	Multiple platforms	7
	R	2
	Mathematica	1
	n.a.	10
Total number of articles		61

General meta-analysis (MA)

A meta-analysis (MA) of RCTs is a statistical method to pool the results of several individual trials in a certain disease and summarize them into a point and its confidence interval (CI) estimate. For these estimates distinction is made between a fixed effect (FE) and a random effects (RE) approach. An FE model assumes that the unknown parameter value is the same for all trials; an RE model assumes that parameter values for the pooled trials follow some distribution. The between-trial variance or heterogeneity (τ^2) has to be estimated for this distribution.

Nowadays, MA methodology is widely implemented [9]. An important issue for the reliability of the results of an MA is the similarity of patients and other trial characteristics across the pooled set of RCTs. Trials can differ in patient-level and in study-level variables. Aiello et al [21]‡ present graphical and analytical tools to identify quantitative criteria to detect these covariate imbalances. These tools are, however, not feasible in an MA with only few trials. Verbeek et al [22]‡ stress that the credibility of an MA depends on the conceptual similarity of the studies and on the statistical heterogeneity.

An MA requires an extensive and complete systematic review (SR) of the medical literature on the disease concerned. RCTs with 'negative' results, i.e. no significant difference between the treatments compared, are less likely to be published, thus leading to publication bias. Publication bias can lead to 'biased' priors for both frequentist and Bayesian analyses and financial disclosure should be a covariate in meta-analyses to prevent investigator bias and assess uncertainty about study effects [23]‡. MA using a FE approach can lead to substantial inflation of the type I errors ([24]‡ and [25]). Both Higgins et al [26]‡ and Chung et al [27]‡ advise against the use of a FE or common model, but also against the testing of homogeneity. They emphasize that the naive presentation of only the mean (μ) of the RE analysis is misleading and estimation of the between-trial variance is just as important as well as its incorporation in a CI for μ. Both for frequentist and Bayesian inference from k RCTs, Higgins et al [26]‡ (as well as Borenstein et al [9]) propose the use of a prediction interval based on a t-distribution with $k-2$ degrees of freedom instead of a Normal distribution to account for the uncertainty in the estimated τ^2. Chung et al [27]‡ discuss the estimation of the between-study variance for small numbers of studies. In such conditions, commonly used estimators

frequently result in a value of 0, thereby underestimating the true heterogeneity. Chung et al, following Borenstein et al [9] and Higgins et al [26]‡, prefer a Bayesian informative prior distribution for the between-study variance based on plausible values from other, similar MAs or on historical data. They propose a Bayes modal estimator and compare its properties to those of other estimators. When study-level covariates are available, meta-regression analysis can be applied to decrease the heterogeneity.

RE models have disadvantages and may add unnecessary complexity to the analysis. To judge whether an RE or linear mixed effects model is appropriate, Demidenko et al [28]‡ propose the RE coefficient of determination, the proportion of conditional variance explained by the heterogeneity of the studies in an MA.

For normally distributed outcomes standard MA theory assumes that variances are known. This theory is often applied to effect sizes with skewed distributions with variances to be estimated. Malloy et al [29]‡ suggest to first apply a variance stabilizing transformation and then estimate point and interval parameters of FE or RE models using stable weights or profile approximate likelihood intervals. Further, a simple t-interval provides very good coverage of an overall effect size without estimation of the heterogeneity [29]‡.

Viechtbauer [30]‡ provides an extensive overview of the capabilities of the 'metafor' package for conducting meta-analyses with R.

Design

Journal guidelines state that a report of an RCT should include a summary of previous research findings, preferable an SR and MA, and explain how the new trial affects this summary. Such a summary should inform critical design issues such as sample size determination [31]‡. Sutton et al [31]‡ stress that the existing evidence-base should be analysed in a detailed way to be able to design future research more efficiently. They propose mixed treatment comparisons (MTC) MA and individual patient data (IPD) MA methods. The contribution of a newly planned RCT to the total evidence is evaluated through its incorporation into an updated MA in various ways. A new trial can be designed and powered in isolation based on the results of a MA or based on the statistical significance of the updated

MA. Heterogeneity between RCTs can seriously influence the power of the updated MA. To better estimate heterogeneity, multiple small new studies can be preferred to a single large study containing the same number of subjects.

Goudie et al [32]‡ found that only few published RCTs reported the use of previous trials to design a future trial and estimate its sample size. They also highlight the importance of adequately considering heterogeneity among studies in an MA, but note that between-study heterogeneity will often be estimated with poor precision. They point out that the process of using evidence from related, but not identical, studies could be formalized by more sophisticated modelling, such as the use of patient-level covariates. Ioannidis and Karassa [33]‡ also emphasize the need to consider breadth, timing and depth of all evidence, including unpublished and on-going studies, for an SR and MA. They consider results from single, early stopped trials unreliable because of chance findings due to multiple testing and inflated treatment effect estimates.

Rotondi and Donner [34]‡ describe estimation of an appropriate sample size for a planned cluster randomized trial by considering the role of the planned trial in a future MA. Sample size estimation can be based on power or reduction in variance or the perspective of non-inferiority. Their approach is based on simulated data using prior distributions for the intra-cluster correlation coefficient, the cluster size and the control event rate. An FE model with dichotomous outcomes is assumed as well as the availability of IPD. They suggest that their method 'may prove particularly useful when dealing with a meta-analysis of a small number of studies'.

Heterogeneity

Between-trial variability or heterogeneity can be tested and estimated. A commonly used measure for heterogeneity between trials pooled in an MA is I^2. Higgins and Thompson [35] derived this measure assuming that all within-trial variances were equal, thus giving all trials the same weight, an assumption that is not met in most MAs. As a better alternative, Wetterslev et al [36]‡ propose a measure of diversity D^2 to describe the reduction in the model variance from an RE MA to an FE MA. They show that $D^2 \geq I^2$ and thus, in general, this will lead to a larger information size, i.e. the required number of participants in an MA.

The derivation of D^2, however, assumes that the FE population average is equal to the RE population average, which requires additional information if this assumption is not met.

Standard meta-analysis methods ignore the uncertainty in the estimation of the heterogeneity parameter [27, 31]‡. Chung et al [27]‡ describe the use of a profile likelihood function (following [37]) to construct a CI for μ or a Wald-type interval based on the observed instead of the expected information.

Turner et al [38]‡ present a method to adjust for differences in rigour (i.e. lack of internal bias) and relevance (i.e. lack of external bias) between studies pooled in an MA. Their bias modelling approach allows decisions to be based on all available evidence with less rigorous or less relevant studies getting smaller weights. Their expectation is that bias adjustment will remove much of the heterogeneity in an MA. Bias adjustment is based, however, on elicited opinions rather than empirical evidence.

Differences in study quality may lead to heterogeneity in findings across studies. Ahn and Becker [39]‡ compared inverse-variance weighting with weights composed from quality scores on the estimated mean effect in an MA. They conclude that quality weighting adds bias in many cases. They prefer to model the effects of components of quality rather than use quality-score weights. Yuan and Little [40]‡ note that the DerSimonian-Laird (DL) estimate for heterogeneity in an RE MA is in general biased when the patient attrition rate depends on the study-specific effect size. Higher completion rates are associated with more extreme effect sizes, i.e. more bias. They propose three methods to correct for this bias, two of which, the re-weighted Bayesian RE model and the Bayesian shared-parameter model work well.

Statistical heterogeneity and small-study effects may affect the validity of an MA. Small-study effects can be seen as a particular case of heterogeneity. To adjust treatment effect estimates for this heterogeneity, Rücker et al [41]‡ introduce the limit MA as a new RE model-based method which leads to shrunken, empirical Bayes estimators. This gives rise to a new measure of heterogeneity, G^2, i.e. the proportion of heterogeneity unexplained after allowance for possible small-study effects in the limit MA.

Rare events

An MA on dichotomous outcome data traditionally pools the summary measures of the individual RCTs (e.g. log(OR) or log(RR)) and their standard errors. This assumes an approximately Normal within-study likelihood with known standard errors, does not account for correlation between the estimate and its standard error and necessitates the use of an (arbitrary) continuity correction in case of zero events. To overcome these drawbacks, Stijnen et al [42]‡ propose an exact likelihood approach leading to a generalized linear mixed model. This approach may be especially advantageous for sparse (event) data.

Lane [43]‡ also notes the limitations of the traditional pooling methods and especially for trials with rare events that are in general not primary outcomes, such as safety outcomes. For these trials, results from an MA should be regarded as only exploratory and hypothesis-generating, especially when there is much heterogeneity between the trials.

Naïve pooling of cumulative proportions of adverse effects can suffer from Simpson's paradox when randomization ratios are not identical across studies. Chuang-Stein and Beltangady [44]‡ discuss three approaches to report these cumulative proportions of safety data. The inverse sample variance weighting is not recommended; Cochran-Mantel-Haenszel weighting and a study size based method produce similar results.

Gruber and Van der Laan [45]‡ compared several estimators of the treatment effect on safety outcomes in an MA for various missingness mechanisms. Their targeted ML estimator is asymptotically efficient and unbiased and has good finite sample performance, also when outcomes are missing at random or missingness is informative. Bennett et al [46]‡ compared the standard Cox PH model to the Firth penalized Cox PH model and to a Bayesian PH model in MAs with survival-type rare event outcome data. They conclude that the Firth model gives less biased estimates of the (log) hazard ratios than the other two models in rare events survival data.

Series of trials

Chambers et al [47]‡ investigated the inclusion of both RCTs and case series in an SR of a rapidly developing technology. Results from non-randomized controlled clinical trials were

also included as case series. The authors found no systematic differences in the primary outcome between RCTs and case series and concluded that the evidence base of an SR can be increased and its credibility strengthened by the inclusion of case series. However, they note some clear drawbacks, such as the absence of a control group and several forms of possible bias. Hee and Stallard [48]‡ propose a hybrid approach to optimally design an entire development plan encompassing phase II and phase III trials by combining Bayesian decision-theoretic elements and frequentist methods. The phase II trials are assumed to be conducted (fully) sequentially with interim decision-making based on a Bayesian cost-utility approach. From the phase II trials, the most promising treatment is identified and evaluated further in a phase III setting. At the design stage, a prior distribution is assumed for the parameters corresponding to the treatment effects for the experimental treatment. The proposed method assumes that the phase II and III trials have the same patient population, primary endpoint and treatment period.

In the context of a rare disease, often the sample size is retrofitted by adapting the desired power and the relevant effect size to the available number of participants. Le Deley et al [49]‡ extended the work of Sposto and Stram [50] to evaluate the efficiency of a series of successive phase III RCTs by performing an extensive simulation study. Parameters for the simulations were, amongst others, the significance level α, the number and size of trials and the effect size; each trial's outcome was of survival type. When the number of available patients is small, results indicate that designs using smaller sample sizes together with relaxed α values yield greater expected survival benefits. The authors assumed that treatment aspects are similar over trials, that many drugs are available for testing and they did not consider interim analyses.

Multivariate outcomes

A multivariate MA of multiple correlated endpoints enables to borrow strength across the endpoints and to calculate joint confidence and prediction intervals [51]‡. When only AD of studies to be pooled are available, an estimate for the correlation between the endpoints within a study is necessary. Riley [51]‡ shows that ignoring this within-study correlation leads to inaccurate pooled estimates in a bivariate RE MA. Only when between-study variation is very large relative to within-study variation, within-study correlation can be ignored. In

general, availability of IPD for all studies to be pooled is desirable. When both IPD and AD are available, a distribution for the correlation can be estimated from the IPD and used as an informative prior distribution for the missing correlations from the AD. Otherwise, sensitivity analyses over a range of values for within-study correlations can be performed. As an alternative, a model with an overall correlation estimate has been proposed by the same author [52].

Jackson et al [53]‡, commentaries [54, 55, 56, 57, 58]‡ and the rejoinder [53]‡ provide a summary of a one day event on 'Multivariate meta-analysis' for the pooling of studies with multiple, often correlated, outcomes of interest. They discuss the multivariate RE model and its assumptions, describe and apply the estimation methods and discuss advantages and limitations of the multivariate MA. The greatest practical difficulty is the estimation of the within- and between-study correlations, for which the authors describe some solutions. The multivariate Normality assumption is often hard to verify as is the linear relationship of the effects between the studies. Multivariate MA can be useful, but also brings complications and issues. One of the commentaries was that a Bayesian approach using prior information in case of few studies with sparse data can be helpful, but will also show the (large) influence of the prior distribution.

Camilli et al [24]‡ compared three multi-level meta-regression models for multiple effect sizes per included study, i.e., a standard multi-level model and an iteratively weighted multi-level model, both with weights based on a Normal approximation to the non-central t distribution, and a multi-level model based on the exact non-central t distribution. The latter model seems to perform better for larger samples. For small samples, however, it is unclear which estimator, a REML- or an MCMC estimator for the between-study variance is better.

Cumulative meta-analysis (CMA)

A CMA evaluates the accumulating evidence of a series of independent RCTs on the same intervention. Its value, amongst others, lies in the early identification of clinical efficacy or harm, thereby discouraging unnecessary future research. However, periodic updating of MAs can inflate the type I error rate substantially and should be accounted for by formal monitoring procedures [31, 59, 60, 61, 62, 63]‡. Borm et al [59]‡ present a rule of thumb that relates the

desired type I error and the P value of the MA to the maximum number of updates. This rule of thumb does not strictly control the type I error, however.

Trends in effect sizes over time can be detected by visual inspection of cumulative plots or by a test of equality of the estimate of the first RCT and the estimate based on the subsequent RCTs or the overall MA. Bagos and Nikolopoulos [64]‡ propose a generalized least squares regression approach to estimate a time trend in effect sizes with a first-order autocorrelation coefficient to adjust for dependence between successive effect size estimates. They applied this exploratory tool in genetic association studies, but also see its usefulness for planning an update of an already published MA. Sutton et al [31]‡ compare two methods to inform prioritization strategies for updating systematic reviews. These methods are only in agreement in case of homogeneity. Although the authors recognize the need to adjust for multiple updating of a CMA, they do not control for this. Herbison et al [65]‡ carried out a number of CMAs to determine the number of trials needed to stable down and get a consistent point estimate. Values for τ^2 and I^2 were no predictors for the number of trials needed nor was the size of the trials. A median of 4 studies were enough to get within 10% of the final point estimate.

Pereira and Ioannidis [66]‡ investigated the occurrence of the "winner's curse phenomenon", i.e. the fact that crossing a significance threshold and at the same time estimating the effect size can result in exaggerated effect size estimates, especially for smaller sample sizes. They evaluated a large number of MAs and found that the magnitude of significant effects is often inflated, but the opposite is also true: if a boundary is not crossed, the estimate may be too small. They argue, following other publications, that CMAs should be adjusted for multiple testing.

Trial sequential analysis (TSA)

TSA combines the a priori calculation of information size for an MA with O'Brien-Fleming monitoring boundaries to evaluate the accumulating trial data and at the same time adjust for the cumulative updating. Calculation of the necessary information size can be performed in various ways, amongst others by adjusting for heterogeneity using I^2 [60, 67]‡. These various information sizes lead to as many sets of trial sequential monitoring boundaries.

Thorlund et al [67]‡ show that the risk of false-positive results and inaccurate effect size estimates can be reduced by the use of TSA. Brok et al [60]‡ find that many published, conclusive MAs are potentially inconclusive when adjusted for the cumulative testing and for heterogeneity. TSA does not allow stopping for futility, however. In a commentary on the previous two papers, Nüesch and Jüni [68]‡ emphasize the need for diagnostic measures (such as funnel plots, stratified analyses and interaction tests) to draw conclusions from an MA. Miladinovic et al [62]‡ recommend to perform and report sensitivity analyses based on acceptable thresholds for the type I error, power and clinically meaningful treatment difference to prevent premature declaration of a significant MA. They note that three MAs prematurely were declared statistically significant, but later turned out to be not. Imberger et al [69]‡ points out that power for two of these three was clearly insufficient to draw a conclusion. Miladinovic et al [63]‡ were the first to apply time-to-event TSA. Like the originally proposed TSAs, they did not control for type II error, which made stopping for futility impossible. As an additional comment they note that application of Bayesian monitoring boundaries may result in narrower credibility intervals. For TSAs with count or time-to-event data, software in R and in STATA is presented and described [70]‡.

Sequential meta-analysis (SMA)

An SMA can be implemented using a triangular test following Whitehead's boundaries approach [71]. With this approach, the type I error and power of a CMA can be guaranteed. Van der Tweel and Bollen [61]‡ compared TSA and SMA by re-analysing a number of published examples incorporating the Paule-Mandel estimator for heterogeneity between trials in the SMA. They showed that for an SMA (1) no prior estimate for total information size is necessary and thus one set of monitoring boundaries suffices; (2) stopping a CMA for futility is an option; (3) the desired power can be specified in the design; (4) point and interval estimates are adjusted for the multiple testing. The estimates for heterogeneity are, however, unstable for a small number of trials. The paper raised some discussion about supposed differences between TSA and SMA [72, 73]‡.

Novianti et al [74]‡ evaluated the properties of estimators of heterogeneity in an SMA. Their simulation study showed that the well-known DL estimator largely underestimates the true value for dichotomous outcomes. They recommend the two-step DL estimator and the

Paule–Mandel estimator for use in an SMA with dichotomous or continuous outcomes.

Prospective meta-analysis (PMA)

A PMA can be designed and executed to combine evidence from new and on-going, similar clinical trials in a prospective way. Its advantages are uniformity of the trial protocol, the intervention, the data collection instruments and the reporting of specific outcomes while allowing individual sites some independence with respect to the conduct of research. The inclusion of several sites increases statistical power to address important clinical questions. In PMA, analysis of pooled results is more facile because of homogeneity of study outcome measures. Besides, IPD enable to conduct stratified analyses and to control for potentially confounding variables. The diversity in study population improves the external validity [75]‡. A PMA is, however, not able to control the generation of new evidence, so the amount, timing and heterogeneity of future trials will not be known in advance. This makes traditional group sequential methods not applicable, but SMA can be applied. Higgins et al [76]‡ propose an informative prior distribution to produce a realistic estimate of the between-trial variance in an early stage of an SMA when only a small number of studies is available. The point estimate is then updated in subsequent stages of the SMA. This semi-Bayes approach incorporates the DL estimator. The false-positive and coverage properties depend on the choice of prior distribution for the between-trial variance. Imberger et al [77]‡ wonder how the parameters for this prior distribution can be interpreted and how heterogeneity is incorporated.

Shuster and Neu [78]‡ argue that prospective group sequential MA methods (such as TSA and SMA) need four essential qualities, i.e. the population effect sizes should be allowed to change over time, independent increments of information from analysis to analysis, robustness against incorrect specification of the information fraction and a physically interpretable effect size. To meet these needs, they impose a separate prior distribution on the effect sizes for each trial, weigh each trial only by sample size and not by the inverse of the variance and apply Pocock's approach to group sequential testing (i.e. a constant nominal type I error probability at each interim analysis). There is no guarantee of power of the PMA, however.

For a recent, practical application of an IPD PMA see Askie et al [79].

Network meta-analysis (NMA)

A single SR or MA of a treatment comparison for a single outcome offers a limited view if there are many treatments or many important outcomes to consider. An umbrella review assembles together several SRs on the same condition. If treatments in the SRs can be connected directly or indirectly in a network, outcomes can be analysed with a multiple-treatment meta-analysis/mixed treatment comparison meta-analysis/network MA (NMA). These analyses can also rank the effectiveness of the treatments in a network, thereby determining the best available treatment. An important issue in an NMA is to examine whether there is incoherence or inconsistency, i.e. whether the effect estimated from indirect comparisons differs from that estimated from direct comparisons. However, the power to detect incoherence is low when there are only a few RCTs. Ioannidis [80]‡ provides key features in the critical reading of umbrella reviews and key considerations for NMA. NMA requires more sophisticated statistical expertise than simple umbrella reviews, but assumes that all data can be analysed together. Most methods for MTMA follow a Bayesian approach. Stijnen et al [42]‡ applied their exact method (see also above under General MA) in an example on NMA. Thorlund and Mills [81]‡ propose flexible methods for estimating the sample size or statistical information and the power in an NMA with both direct and indirect treatment comparisons. Their sample size formulas correct for heterogeneity using I^2.

To assess the effect of a particular combination of drugs, Thorlund and Mills [82]‡ propose an NMA model with an additive-effect parameter. Such a model gains precision by assuming full additivity of treatment effects, that is: when the effect of the treatment combination is equal to the sum of the stand-alone effects. The additive-effects model is superior to the conventional NMA model when full additivity holds. The two models are comparably advantageous (in terms of a bias-precision trade-off) when additivity is mildly violated. When additivity is strongly violated, the additive effects model is statistically inferior. When additivity can be assumed, it seems reasonable to prefer the additive effects MTCMA model above the conventional model.

An NMA assumes similarity across the pooled set of trials in terms of patient population and trial characteristics. Naci and O'Connor [83]‡ describe the possible benefits of a prospective NMA, such as access to IPD by regulatory agency statisticians, to evaluate comparative

efficacy and safety of more than two drugs. Information from both direct and indirect comparisons from a network of trials can provide (far) more information, especially on safety, than just pairwise MA. They urge researchers, manufacturers and regulators to collaborate on future trial designs and analyses. Regulators having access to IPD could also help to inform patients more completely about new treatments. They note, however, that FDA and EMA might not be allowed to use proprietary information from the marketing application of one drug in the evaluation of another.

Bafeta et al [84]‡ performed a methodological review of reports of NMAs. They conclude that essential methodological components of the review process, like conducting a literature search and assessing risk of bias of individual studies, are frequently lacking in the reports. They call for guidelines to improve the quality of reporting and conduct of NMAs. We refer to the reader to Chapter 2b and Chapter 7 for a more elaborated discussion on NMA.

Aggregate data (AD) vs individual patient data (IPD)

Traditionally, an MA combines evidence from related RCTs based on aggregate study-level data. Increasingly, IPD are used. Riley et al [85]‡ go into the rationale behind IPD MAs. IPD are not needed if the required AD can be obtained in full from publications. However, IPD MAs are potentially more reliable than AD MAs. Use of IPD can increase the power to detect a differential treatment effect, allows adjustment for covariates on patient-level instead of study-level and is particularly advantageous for time-to-event data. A disadvantage is that the IPD approach can take lots of time and costs, and often requires advanced statistical expertise (like FE and RE MA) to preserve the clustering of patients within studies. Increasing use of PMA on IPD is advocated.

To identify a possible source of treatment effect heterogeneity, a treatment-covariate interaction (with the covariate defining the subgroups of interest) can be estimated from a regression analysis on IPD. Kovalchik [86]‡ presents an AD EM-algorithm that is equivalent to the ML estimates for an IPD linear RE MA with a patient-level treatment-covariate interaction term for a categorical covariate, when the model's variance parameters are known. The presented methodology does not replace an IPD MA, but provides a good AD approximation to a specific kind of IPD interaction model when patient-level data cannot be obtained.

Discussion

Research in rare diseases faces two problems. First, a small number of participants available per trial, and second, usually only a small number of trials targeting the same (new) treatment is possible. In this review we described statistical methods to combine results of series of trials, as published in a recent period of five years. Various search engines were explored. This is specifically important for a methodological review. For example, with the extension to Scopus, 39 unique papers were identified. In total, 61 papers were included in this review. We categorized the relevant methodology according to the type of (meta-)analysis and assessed its usefulness and limitations in small populations. The focus of this review is on methodology. Completeness is less of an issue in methodological research. A more extensive search could identify additional papers, but is unlikely to provide new insights. In other words, our search will reach a stage ('theoretical saturation') where identifying more articles will not render further methodological perspectives [87].

In general, an MA is a well-accepted way of pooling results from a series of trials. Various approaches to MA have been described and evaluated in the past. Herbison et al [65]‡ concluded that a median number of 4 studies are needed to get within 10% of the final pooled point estimate, where they based this final value on a minimum of 10 trials and assumed it the true value. They restricted themselves to FE estimates based on 95% CIs and did not adjust the CIs for multiple testing. They recognize that it is impossible to predict which SRs with a small number of studies will be correct in the long run.

Simulation studies with survival type outcomes showed that designs using smaller sample sizes and relaxed α values yield greater expected survival benefits than traditional design strategies that aimed to detect a small difference with high level of evidence [49]‡ with reference to Sposto and Stram [50]. These studies focused on personalized medicine, but can also be useful for RCTs in rare diseases. Research has to confirm the results for dichotomous and continuous outcomes. O'Connor and Hemmings [18] also suggested relaxation of the type I error.

Both Miladinovic et al [62]‡ and Nüesch and Jüni [68]‡ cite Egger and Davey Smith [88] that *'results of meta-analyses that are exclusively based on small trials should be distrusted - even if the*

combined effect is statistically highly significant. Several medium-sized trials of high quality seem necessary to render results trustworthy.' This citation is opposite to the suggestion by IntHout et al [89] that '*evidence of efficacy based on a series of smaller trials may lower the error rates compared with a single well-powered trial*'.

Most research in MA acknowledges the need to incorporate heterogeneity into the effect point- and interval estimates. The properties of the estimators are not well-known though for a small number of trials. Various authors note that both I^2 and τ^2 as measures for heterogeneity can be unreliable and unstable in an MA with a small number of trials. Estimating heterogeneity is considered more important than testing it. To account for the uncertainty in the estimated value of τ^2 in a CI for the pooled effect size, the use of a t-distribution with $k-1$ or $k-2$ degrees of freedom (with k the number of trials pooled) instead of a Normal distribution in a CI for the pooled effect size is proposed [26, 29, 62, 68]‡. This implies that the number of RCTs to be pooled in an MA should be at least three.

For continuous outcomes a variance stabilizing transformation is advised [29]‡ before estimating the confidence interval. The method-of-moments estimator according to DerSimonian and Laird (DL) to estimate the between-study heterogeneity parameter τ^2 is widely applied ([9] and [28, 29, 38]‡) and is also standard in software such as Review Manager [90] and Comprehensive Meta-analysis [91]. Yuan and Little [40]‡ observe a bias in the DL estimator leading to too narrow CIs. Turner et al [38]‡ and Novianti et al [74]‡ note that alternative estimators such as proposed by DerSimonian and Kacker [92] might be preferred. Their properties, and those of other recently proposed estimators [16], in a small number of RCTs have to be explored.

Case series of the use of therapeutic procedures or devices can be included to strengthen the evidence in an SR, although Chambers et al [30] mention some drawbacks. The contribution of case series of drug use for an SR and MA in rare diseases has to be further explored. Hee and Stallard [48] propose an optimal decision-theoretic design of a series of phase II clinical trials followed by a phase III RCT. Their approach is a hybrid one, in that it assumes prior distributions for the success probabilities in the phase II trials, followed by a classical frequentist hypothesis test. This proposal can be useful in rare diseases, but its application in RCTs with non-dichotomous outcomes has to be investigated further.

Frequently, an MA is updated with results of one or more newly published RCTs, leading to a so-called CMA. In general, such an update does not control for multiple testing, thereby risking an increase in the overall type I error. A TSA or SMA design, on the contrary, guarantees the overall type I error. The use of a TSA, an SMA or a PMA enables to stop a series of trials for efficacy or futility, thereby leading to efficiency gains and thus ethical and/or economic benefits. Ideally, TSA should be applied prospectively with clinically relevant pre-specified treatment differences, type I and type II errors [61]‡. These authors also see a role for sensitivity analyses.

Thorlund and Mills [82]‡ use I^2 to correct for heterogeneity in an NMA. Its use as a measure of heterogeneity can, however, be debated. It is, for example, known to increase with the number of patients included in the studies in a MA [93]. Wetterslev et al [36]‡ conclude that their proposed measure D^2 seems a better alternative for trial diversity and for adjustment of the required information size. Demidenko et al [28]‡ developed a coefficient of determination to measure the strength of the presence of random effects in a model. It is unclear what its additional value is to I^2. Higgins et al [76]‡ consider clinical research a sequential process where SMA can play a role in the design of a new trial, since the amount of further information that would be required can be determined. They notice, however, also some points of attention. One is whether or not a correction for multiple looks to cumulative data is needed. Another is the poor estimation of τ^2 from a small number of studies. Then realistic prior information is necessary, but the choice of the prior distribution is crucial in the early stages of an SMA. Undertaking an MA in a fully Bayesian way has the advantage that no correction for multiple looks is necessary for inference, but frequentist properties, such as type I errors, can be inflated.

Rücker et al [41]‡ suggest to adjust treatment effect estimates for small-study effects, leading to shrunken, empirical Bayes estimates. These estimates are approximately unbiased when the number of trials in an MA is at least 10. The approach depends on the estimator for τ^2, which was the DL estimator, which is known to underestimate $\tau^2$for dichotomous outcomes. The remaining amount of heterogeneity, termed G^2, varies considerably depending on the estimator used. Further research will be needed to investigate whether this approach is useful for MAs with less than 10 RCTs and with other estimators.

Especially in rare diseases, multiple outcomes will (have to) be examined simultaneously. In that case a multivariate MA as proposed by Jackson et al [53, 54, 55, 56, 58, 57]‡ may show potential, but also raises concerns. In particular, the statistical properties for a small number of small samples, imprecise between-study (co)variances, unavailable within-study correlation estimates, a possible large number of parameters to be estimated, and missing outcomes in some but not all trials require further study. It also makes clear that IPD will have to be available for RCTs in such MAs. Riley [94]‡ also points to the important role of a multivariate MA in evidence-based decision making. His approach assumes the within-study correlation is given and known, though. Comparison of this approach with an earlier proposed alternative [52], a model with an overall correlation estimate, in small populations deserves further investigation. Stijnen et al [42]‡ presented an extension of their exact likelihood method for dichotomous outcomes into a multivariate MA. Their model can also be applied with rare event outcomes.

Both frequentist and Bayesian approaches are applied to combine successfully the extracted data from several trials. Their application in the field of rare diseases is one possible way to sufficiently support a treatment effect. Measures for heterogeneity can be unreliable and unstable in an MA based on a small number of trials. An option is to formulate an informative prior distribution around τ^2. This prior can be updated in an SMA with the result of a new MA leading to a posterior distribution, which in turn forms a new prior. However, for small data, Bayesian posterior probabilities may depend heavily on the choice of the prior distribution. Higgins et al [76]‡ prefer a Bayesian approach, especially for prediction. They note, however, that their approach does not lend itself well to rare events. Furthermore, it is not clear that strict control over false-positive findings is important in this context, since a small, non-statistically significant, signal should still be investigated when the adverse effect is major. Both Higgins et al [26]‡ and Chung et al [27]‡ prefer a Bayesian informative prior distribution for the heterogeneity parameter. The proposed Bayes modal estimator prevents zero (i.e. boundary) estimates and shows good properties for a small number of studies. Most methods for NMA follow a Bayesian approach. Ioannidis [80]‡ notes that the power to detect inconsistency in an NMA is low when the network consists of only a few (small) trials.

In general, availability of IPD for all studies to be pooled is desirable, particularly in rare

diseases. There, the role of regulators is a further point of attention, because of the remark made by Naci and O'Connor [83]‡ that FDA and EMA might not be allowed to use proprietary information from the marketing application of one drug in the evaluation of another.

Recently, several initiatives have been started to facilitate and promote the sharing of clinical trial data. Members of three EU-FP7 projects on small populations (**Asterix**, **Ideal** and **Inspire**) together with representatives from regulatory agencies, scientific journals and industry addressed the arising intricate biostatistical questions such as the interpretation of multiple statistical analyses, both prospective and retrospective as well as the issue of data protection which is most prominent in the setting of rare diseases [95].

Conclusions

For evidence-based decision-making on a (small) series of trials in small populations, our review has led to several directions for further investigation: 1) frequentist properties of estimators for heterogeneity between trials; 2) use of exact (likelihood) methods; 3) value of prospective meta-analysis in drug development; 4) combination of observational, historical and trial data to ensure that every patient contributes as much information as possible; [18, 96] 5) relax the type I error probability; 6) focus on multiple outcomes per patient; 7) combination of IPD with AD; 8) special attention for the evaluation of rare events, such as safety outcomes.

Acknowledgements

The authors would like to thank Stavros Nikolakopoulos for helpful comments on the manuscript.

Chapter 2b

Bayesian evidence synthesis for combining results of series of a few small trials

K Pateras

L Spineli

KCB Roes

Thesis exclusive

Abstract

Classical and Network meta-analysis can play an important role in clinical research for rare diseases, where it is difficult to conduct large randomized clinical trials and there is a large unmet need for new treatments. This review aims to summarize Bayesian methodology for meta-analysis specifically for small populations and to provide directions for application in clinical drug development for rare diseases. We conducted a 9-year scoping review and identified methodologies applicable for combining results for a series of a few small available trials, excluding variations methods for the design of a new trial. We summarized methodology divided in methodological domains of pairwise and network meta-analysis. Secondly, by utilizing selected European Public Assessment Reports of approved drugs with an orphan designation, we assessed Bayesian meta-analysis methods for application in drug development of rare diseases. Only a few articles dealt with series of trials in small populations directly and most of these focused on pairwise meta-analysis. Limited attention is paid to adapt standard asymptotically valid approaches to the finite sample case of a small number of small trials. Relevant methods facilitate the inclusion of data from prior trials or meta-analyses through prior distributions for parameters that cannot be reliably estimated (i.e. between-trial variance τ^2). Our assessment of approved orphan drugs indicated that non-zero between-trial variance can occur, even if these trials have identical design. Our review did not identify clearly methods applicable for this particular setting. Bayesian meta-analysis methods may overcome methodological difficulties that are inherent to evidence scarcity. Nevertheless, currently available Bayesian meta-analysis methods tailored to small populations are not common. Marketing authorization of orphan drugs could benefit from Bayesian methods in the context of series of small trials but for the proper application of Bayesian meta-analysis in an orphan drug evaluation more methodological developments are needed.

Background

Randomized controlled clinical trials (RCTs) are considered the gold standard for comparing and evaluating the efficacy of medical interventions [97, 98]. Considering the number of people affected by a rare disease (prevalence between 1/2500 and 1/1000 [1]), in general a sufficiently sized RCT is not always feasible, while smaller trials produce underpowered and non-conclusive evidence [5, 99]. When evaluating a novel intervention in rare diseases (orphan drug) the number of available trials is small; usually the number of participants per trial is limited as well [100]. Similarly to Chapter 2a, please note that the list of references is split into two parts: 1) references to other publications and 2) references to papers in the review ([‡]).

In 2006, the guideline on clinical trials in small populations was published and referred to Bayesian methods as potentially "*advantageous when faced with small datasets, although introducing prior beliefs is often a concern in drug regulation*" [19]. Specific features of Bayesian methods (i.e. flexibility, incorporation of external evidence, easy implementation of complex models) have led to the increasing use of this statistical framework, particularly for phase I trials and adaptive designs [101]. Still, Bayesian methods are not established in confirmatory trials yet and their use is often debated [19, 102].

Bayesian approaches for design and analysis of clinical trials in rare diseases have been advocated by scholars extensively, as they can maximize information from a limited number of subjects by combining external information with trial evidence [103, 104]. For example, in paediatric trials, clinicians face the challenge of extrapolating results from adult trials to a paediatric population. Bayesian approaches may be used to increase the efficiency of the paediatric trial by borrowing strength from adult trial or series of trials [105, 106].

Several statistical methods for combining a series of trial results have been developed. These methods infer on an intervention's effectiveness or generate hypotheses for the planning of new trials, among other purposes [107, 108]. Meta-analysis (MA), first defined by Glass in the social science literature as *"the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings"* [109], is established as a statistical tool to integrate the results of several individual studies on two interventions. MA benefits include

the possible increased power and/or the added generalizability of results in comparison to a single study [9, 110].

According to international guidelines [102], drug licensing in large scale diseases requires two pivotal phase III trials to be successful as part of the results that are submitted for a market authorization application. Such a requirement may not always apply for rare diseases and conditional approval is granted, in case of a large unmet clinical need [19]. Nevertheless, in rare diseases a number of controlled or uncontrolled phase II trials often may exist, prior to conducting a confirmatory phase III trial (i.e. [111, 112, 113]). Thus, a MA of these clinical studies as part of a drug development plan for a rare disease deserves attention.

The classical pairwise meta-analysis (PMA) is restricted to comparing two interventions at a time, which would typically suffice in a regular drug development plan. In rare diseases, an increasing collaboration between sponsors (industry or academia) occurs, allowing for the evaluation of several treatment options in parallel. Network meta-analysis (NMA), an extension of classical PMA, evaluates the relative effectiveness of multiple interventions by synthesizing the available evidence across a network of studies that compare different sets of interventions. The ultimate goal of NMA is to provide a coherent ranking of interventions and assist in decision-making. As in rare diseases the restricted number of patients prohibits the conduct of a large number of RCTs, NMA may act as a bridge over evaluating multiple intervention options; out of those options, all treatments need not be directly compared in every RCT in the NMA.

The small sample restriction of rare diseases may make it more acceptable to consider Bayesian methods for classical PMA and NMA in a regulatory setting [19]. Bayesian techniques are used to facilitate the interpretation of results, since they allow for probabilistic statements on the effectiveness of the compared interventions [114]. In addition, Bayesian methods in MA might be able to tackle issues that are enhanced in small populations (i.e the proper synthesis of few number of small trials [13][‡], the possibly problematic crude synthesis of adverse rare events [115]) by incorporating historical evidence in the form of prior distributions.

With this background, we assessed current Bayesian MA and NMA approaches in the context

of new (pharmaceutical) treatments development for rare diseases. Thus, this article reviews the methodology, draws examples from the European Medical Agency's Public Assessment Reports, and reflects on issues relevant to the application of Bayesian MA methodology in drug development of rare diseases. First, we describe two general meta-analytical areas: (1) pairwise meta-analysis (PMA), which includes multiple outcome meta-analysis (MMA), and (2) network or mixed treatments meta-analysis (NMA) and we refer to them under the term "meta-analysis" in the remainder of this paper. Next, we focus on aspects (domains) of meta-analysis that demand attention in case of orphan drug development: (3) heterogeneity – τ^2, (4) individual patient data (IPD) and (5) reporting and trial design biases and (6) rare events.

The article is organized as follows. We describe the search strategy, introduce the scope of each meta-analytical area and summarize the eligible methodological articles as retrieved from our literature review. Subsequently, we summarize available statistical approaches in four domains that are particularly important for rare diseases. We reflect on our collected methods and their suitability through a pragmatic evaluation of typical examples from rare disease drug development. We conclude with a discussion and provide recommendations for practice.

Methods

Search strategy

We conducted a scoping review of research published from 1.1.2009 until 30.12.2017. In order to include articles from journals that were currently indexed in only a specific search engine we performed a broad search including five search engines, namely, PubMed, Web of Science (WOS), Scopus, JSTOR, and Cochrane collaboration Library (Supplementary material 1). A brief description of the overlap among the 5 databases is provided in Supplementary material 1 (Figure 1).

We remained liberal at including articles in each step to avoid omitting important articles. We included three main keywords and variations at each search, "trial", "meta-analysis" and "Bayes". The only keywords that were used to limit the number of articles were, "a meta-analysis", "a systematic review", "Phase I" and "phase IV". Initially, a range of keywords referring to small populations was utilized as an additional search term (i.e. rare diseases,

small populations, few trials). However, we decided to exclude this term, as the number of resulting articles was very limited. Once we were close to completing the review, we updated the search by monitoring journals of methodological interest on the topic prospectively (Supplementary material 1).

Eligible methodological articles

To exclude papers on applications of classical and network MA, we eliminated most clinical-focused journals (a list of the journals we considered can be found in Supplementary material 1 (Table A2)). Then we excluded search entries if they: (1) were published in a different language than English, (2) were included in conference posters or proceedings, (3) were books / chapters, (4) consisted of applications of meta-analyses that did not have a specific methodological interest (5) had no relevance to meta-analysis, (6) described meta-analysis in a non "clinical trial" context (i.e. genetics), (7) were discussing n-of-1 trials, (8) if their full text was not available or (9) were utilizing a meta-analysis mostly for the design of a future trial.

Selection of methodological articles

One reviewer (KP) judged the initial selection of articles by title and excluded articles given the above characteristics. Then, two reviewers (KP, LMS) independently judged the remaining articles by title and abstract to (I) include, (PI) probably include, (PE) probably exclude or (E) exclude. Articles with a concordant score from both reviewers were either included (I or PI) or excluded (PE or E). The discordant articles were discussed with a third -independent- colleague (KR) until consensus was reached. Finally, following a similar procedure, the remaining articles were evaluated on the basis of their full text (Figure 2b.1).

Data extraction

Concerning the eligible methodological articles, we extracted information on the meta-analytical area (i.e. PMA/MMA or NMA), the year of publication, the statistical software used, the types of outcome (i.e. binary, continuous, time-to-event) and, if applicable, their specific methodological components (i.e. heterogeneity, individual patient data).

European Public Assesment Reports (EPARs)

Finally, we searched EPARs of the European Medicines Agency and selected specific examples of approved orphan drugs to demonstrate conditions for which a Bayesian meta-analysis could be utilized. We considered EPARs of approved orphan drugs, published from 2006 until today. Our selection of examples represents a range of rare conditions with different characteristics [100].

Results

The search resulted in 15,767 articles, including duplicates, out of which 2,582 from PubMed, 5,449 from Scopus, 5,252 from WOS, 2,436 from JSTOR and 48 from the Cochrane Library (Figure 2b.1). After evaluating the full text of 97 articles, in total 31 eligible articles were included (Figure 2b.1). Table 2b.1 provides their descriptive characteristics. Most articles concentrate on dichotomous outcomes. More than half describe methods of PMA via the use of R or WinBUGS. NMA is mostly implemented via Bayesian hierarchical models [116], which explains the use of WinBUGS / JAGS program by half of the articles. Since MAs that utilize IPD are less prevalent due to the limited availability of data, as expected, the vast majority of articles explores techniques that use (summary) aggregated data (AD). We provide an analytical table of the 31 eligible articles, alongside their characteristics in Supplementary material 2 (Table A1).

Pairwise univariate meta-analysis (PMA) - Multiple outcome meta-analysis (MMA)

The frequently used Bayesian random-effects (RE) PMA model has a two-level hierarchical structure [117]. In Bayesian inference all unknown parameters can be considered random variables and need a prior distribution. The choice of prior distribution for the overall treatment effect is not trivial. Usually a diffuse normal prior is placed on the overall treatment effect among other choices [118]. However, since the choice of prior for the between-study variance (heterogeneity - τ^2) parameter may impact the posterior inferences, it remains a controversial topic in a PMA of a few small studies [26].

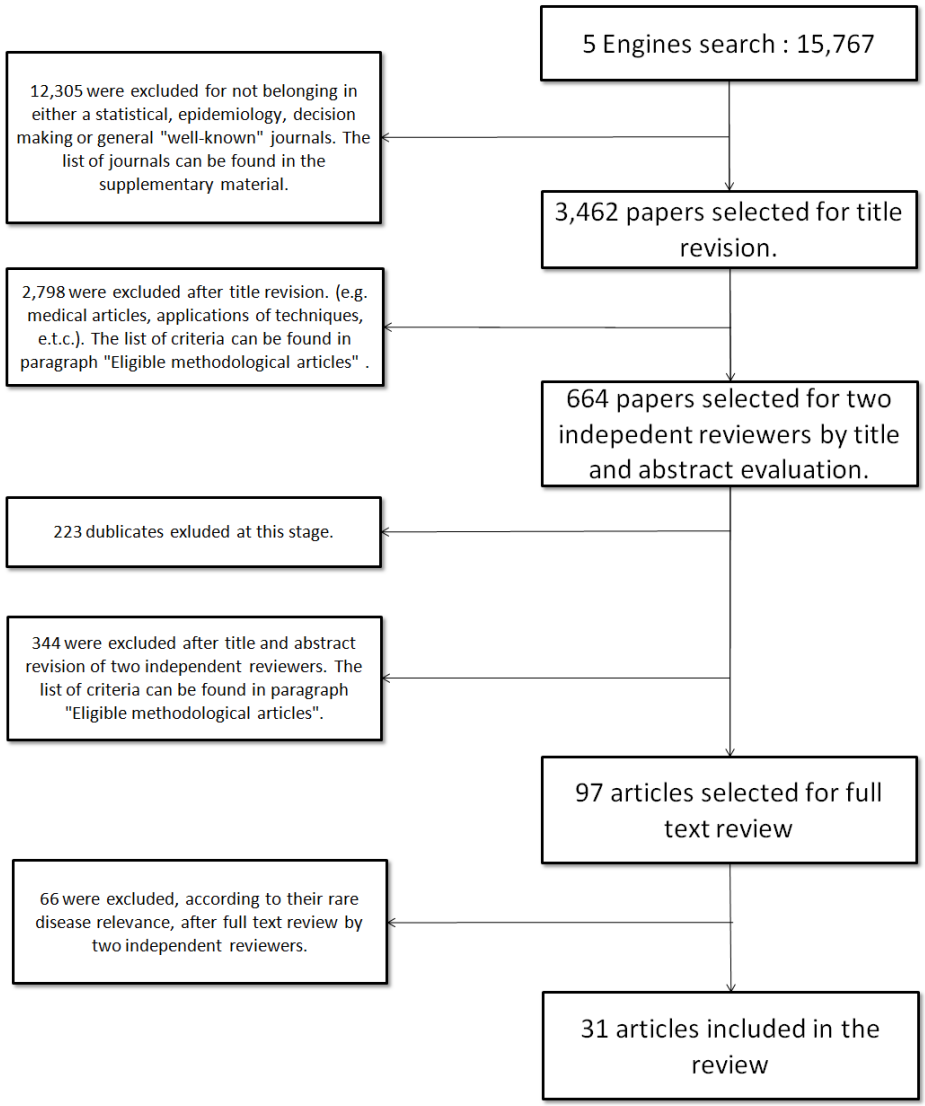

Figure 2b.1: Chapter 2b - Flow diagram of the search strategy

Often in RCTs more than one outcome exists (i.e. improvement of symptoms, treatment discontinuation). When we suspect these outcomes to be correlated, employing an MMA is recommended, in order to account for their correlation. Separate PMAs of correlated

Table 2b.1: Characteristics of the articles included in the review. *MA = meta-analysis, PMA = Pairwise meta-analysis, NMA = network meta-analysis, MTC = mixed treatment comparison, AD = aggregated data, IPD = individual patient data, n.s. = not applicable / not specified.

Characteristic	Category	Count #
Year of publication	2009	2
	2010	1
	2011	2
	2012	3
	2013	8
	2014	3
	2015	2
	2016	7
	2017	3
Input data	AD	24
	IPD	3
	AD+IPD	4
Type of outcome	Dichotomous	17
	Continuous	1
	Time-to-event	1
	Multiple	12
Type of analysis	Univariate	27
	Multivariate	4
Type of MA	PMA	23
	NMA	8
Model approach	Bayes	22
	Frequentist and Bayes	9
Program	WinBugs/Jags	11
	Multiple platforms	8
	R	4
	Mathematica	2
	n.s.	6
Total number of articles		31

outcomes may lead to overestimated variances and biased estimates [94]. Compared to a PMA, an MMA analyses multiple outcomes simultaneously [119]. A simple RE MMA model assumes a multivariate normal distribution for the multiple outcomes and includes a covariance matrix that consists of within- and between-study variances and correlations.

In rare diseases, the number of studies remains small, and usually not enough information becomes available for estimating the variance-covariance matrix correctly [120]‡. For example, Wei et al discuss that at least 15 studies may be needed to analyse their unstructured bivariate model robustly [120]‡. Due to the limited number of studies, prior considerations might highly impact the posterior estimates. In such a multiple outcome setting, as shown in the

literature, the Bayesian approach may be particularly effective when synthesizing six or fewer studies ([121] and [122]‡) via the use of informative priors on the between-study variance and correlation parameters ([123] and [124, 125]‡).

MMA methods exist that simultaneously analyse outcomes, which are not reported by all incorporated studies [120]‡. In addition, methods that are capable of combining different types of outcome (i.e. continuous, dichotomous) could be beneficial in rare diseases [126]‡. The success of such approaches rests in their ability to adjust for the missing outcomes by borrowing information from the other studies. Nonetheless, when the number of outcomes is relatively larger than the number of available studies, estimating the between-study correlation becomes challenging and additional assumptions on the variance-covariance matrix are commonly applied (i.e. homogeneity among within-study variances and/or within-study correlations) [120]‡. For the special case of two correlated outcomes, a Bayesian MMA can be employed without accounting explicitly for the within-study correlation, by directly modelling their overall correlation and their within-study variances [52]. This approach may become problematic in cases where the within-study variability becomes relatively large [52]

Network meta-analysis (NMA)

NMA (also known as multiple treatment meta-analysis or mixed-treatment comparison) is considered an extension of PMA [127, 128] and can be further extended to multiple outcome NMA [10, 129]. NMA allows inferences to be made on interventions which are not directly (head-to-head) compared in any trial, by examining the relative effects of these interventions against (at least) a common comparator; hence, indirect evidence is produced. The validity of the indirect evidence depends on the transitivity assumption which states that there are no differences between directly observed and indirectly observed intervention effects beyond the between-trial variance and consequently, any missing intervention in each trial is missing at random [10]. The statistical manifestation of transitivity is known as consistency and refers to the agreement between direct and indirect evidence (usually derived by more than one routes) in a closed loop of evidence that comprises a "circuit" of connected interventions. If consistency holds, then direct and indirect evidence can be pooled to obtain a mixed estimate of a comparison. Salanti provides a description of transitivity and consistency in a more-

detailed, fruitful discussion. [10].

Both a frequentist and a Bayesian perspective can be applied to NMA. Bayesian techniques in NMA seem to dominate during the recent years, due to their ability to handle complicated models and support probabilistic inferences, such as the ranking of competing interventions [116, 129]. Nevertheless, the availability of only a few trials may negatively impact the credibility of the NMA results, as elements of the covariance matrix become inestimable [130]‡. Diffuse priors are usually placed on location and dispersion parameters, however, weakly informative [131] and informative empirical priors [132] on the covariance matrix have been recommended when the available evidence is sparse [133, 134]. In so sparse conditions, NMA performs a ”borrowing of strength“ across both trials and the assumed NMA structure [135]. Salanti et al [114] highlighted that ranking probabilities of the treatment effects are prone to small changes of the posterior in case of a small number of trials, and therefore the estimated ranks (and their uncertainty) should be reported, for example, together with the effect size of each intervention relative to a reference treatment.

In rare diseases, networks of trials are poorly connected. Closed loops of intervention are particularly scarce and indirect evidence dominates, rendering a thorough examination of the consistency assumption and a formal synthesis of the evidence network challenging, if not impossible ([136, 137] and [138]‡). Nonetheless, when facing a few studies, a Bayesian NMA with the use of informative priors on the between-study variance might offer an alternative over the unstable inferences of frequentist methods for detecting inconsistency [139]. A Bayesian synthesis enables a more reliable estimation of heterogeneity [13]‡ and since extent of heterogeneity and the likelihood to detect inconsistency are inversely related, the latter can be identified more robustly [10]. In general, when less than five trials are synthesized or when the network is poorly connected, the performance and the reliability of NMA deteriorate [140, 141].

A fully Bayesian arm-based model has been developed to detect inconsistency in an NMA of binary outcomes by the use of discrepancy factors [142]‡. This approach showed the ability to handle sparse data more efficiently, in comparison to the Lu and Ades contrast-based model [133].

Heterogeneity – τ^2

In the presence of a few trials, the accurate estimation of statistical heterogeneity becomes challenging. Inaccurate estimation of heterogeneity may compromise the quality of inferences obtained from the previously mentioned meta-analytical domains and result in unreliable inferences. In this case, the use of plausible values for τ or the application of a fixed-effect model has been suggested [26]. The latter assumes that studies estimate the same true effect. This assumption may be realistic, if the main objective would be either to demonstrate a treatment effect at least for a specific group and no unexplained heterogeneity exist or to perform hypothesis testing [143]. However, the fixed-effect assumption becomes more difficult to defend, i.e. if the main objective would be to estimate a treatment effect in a broader population [9].

In order to remain objective and make data-driven inferences, vague or low informative priors can be assigned on the heterogeneity. An extensive simulation study evaluated operational characteristics of various priors on heterogeneity, based on vague priors for the heterogeneity parameter. It showed that inference for the treatment effect can become unstable with relatively sparse data (less than five trials and less than 100 patients per trial), therefore, sensitivity analysis on the prior selection for the heterogeneity parameter should always be considered [144]. The selection of the prior distribution for heterogeneity impacts the prediction of the treatment effect in a future relevant clinical trial as well [145]‡. The half-t family of priors is introduced as a valid and robust alternative when the available evidence is scarce [146]. The half-normal prior, a member of this family, has been evaluated and is suggested for a MA of a few small trials in comparison to standard frequentist alternatives ([144] and [13, 147]‡). In such sparse conditions, plausible ranges for the heterogeneity priors are usually suggested to aid inferences and to provide meaningful results [131]. Finally, one could apply a reference prior which has the ability to maximize the data impact on inference [148]‡ . Currently, this prior applies only to the basic normal-normal hierarchical model which appeared to provide improper coverage in sparse conditions (i.e. zero events) under a number of alternative hypotheses [149, 150].

As an alternate strategy for estimating heterogeneity, the idea of using historical MAs from the same therapeutic area to inform a Bayesian MA was introduced by Higgins and Whitehead

[135]. Recently, various attempts to summarize knowledge on existing MAs have been initiated [151] and [152, 153, 154]‡). Turner et al and Rhodes et al used a large database of Cochrane reviews in order to provide predictive distributions for the heterogeneity parameter tailored to different medical settings depending on the nature of the outcomes (i.e. subjective) and comparisons (i.e. pharmacological versus placebo) ([151] and [153, 155, 156]‡).

In the context of treatment comparisons with sparse data in NMA, authors advocated against the use of the standard homogeneous model and they suggested the use of informative variance priors instead [157]‡. Even under a frequentist pooling of study-specific effects, heterogeneity estimators derived through Bayesian theory have been suggested as promising, especially in the context of a few trials ([27] and [158, 159]‡).

Individual patient data

The use of IPD is regarded as the gold standard for performing a MA and is a special topic of current research interest [160, 161]. Clinical and methodological sources of heterogeneity can be best explored when having access to IPD by one-step (standard general linear regression) or two-step (meta-regression) and subgroup analysis. In addition, several advantages, such as the report of lower absolute biases, have been argued when conducting a Bayesian IPD MA of survival endpoints [162]‡.

Often IPD are not widely available and as a result, methods that combine AD and IPD for a subset of trials have been developed in the PMA [163] and NMA context ([164] and [165, 166]‡). These methods are regarded as a way of creating precise estimates of treatment effects and evaluating in depth sources of heterogeneity while eliminating the risk of ecological bias observed in aggregated data. Although the sole use of AD data may be misleading [164], NMA models that combine AD with IPD run the risk of becoming unstable when the number of contributing trials per comparison is limited [166]‡. In such cases, specific a-priori assumptions, such as a common interaction or exchangeable coefficients may provide a gain in precision or allow for borrowing of strength across the trials [166]‡.

Reporting and trial design biases

Biases can be introduced in a MA through the inclusion of studies with methodological and reporting limitations. Inadequacies in the design characteristics of the contributing trials (i.e. allocation concealment, missing data) and deficiencies on the reporting of results may compromise the internal and external validity of an MA. A partially subjective resort to deal with variation in the quality of the studies would be weighting, by using empirically-based priors [167]‡. Studies are divided into low and high risk bias, based on a specific bias domain (i.e. adequate or inadequate allocation concealment) and they are entered in the model. Unless information from the low risk bias studies is really limited, the weighted inclusion of high risk bias studies will provide only a small gain in precision [167]‡.

Publication bias is the most prominent type of bias in MA and has received attention from clinicians and methodologists since the 1970s. Since in our context, studies are usually small, non-significant results may be easily subject to the *"file drawer problem"* (studies that do not support the hypotheses of researchers often end up in the researchers' file drawers) and the risk of publication bias may be high in small population MAs. Selection models, such as [168], may address missing studies more efficiently using probabilistic statements within a sensitivity analysis framework, as opposed to full meta-regressions when applied in rare diseases.

The Bayesian framework offers the flexibility to incorporate data from different study designs through a three-level hierarchical model [169]‡. Such a model accounts not only for the between-study variance within each design setting, but also for the between-design variance, namely different information per trial design, resulting in possible reduction of *selection biases*. Nonetheless, usually such multilevel models may become unidentifiable for a few small trials.

Often, individual trials with common endpoints have dissimilar treatment arms [170] or report different effect measures (i.e. hazard ratios in some trials whereas odds ratios in others) [171]‡. Methods that combine different type of data sources (i.e. variety of study designs, variety of study effect measures) in order to minimize *bias due to outcome restrictions* are an option for a MA of a few trials as well ([170] and [171]‡).

Rare events

Despite its flexibility, the standard normal Bayesian hierarchical model may perform sub-optimally when applied on dichotomous outcomes, especially in sparse conditions that introduce zero events. Zero events can highly impact the estimation of both the overall treatment effect and the heterogeneity. Robust methods tailored to deal with that issue via the use of exact distributions have been introduced in a Bayesian framework ([172]‡ and [173]‡). For example, Moreno et al refrain from using continuity corrections by utilizing exact binomial distributions and suggest alternatives to the standard normal linking distribution between the overall effect and the study-specific effects [172]‡.

Vazquez et al deviate from the common normal approximations and apply an automatic sensitivity analysis to hyperparameters of the prior distributions [174]‡. Even though their method can be readily extended to sparse and disconnected networks, it has not been compared with other models and specifically in situations of small populations. Another methodology to deal with rarity of events, the so-called B-Bird method, has been introduced by Tang et al [175]‡. Their method has been designed for the risk difference and utilizes historical information available on the rarity of events via the manipulation of the hyperparameters of a Beta prior on a binomial likelihood. It is shown to outperform the classical risk difference model of Warn et al [176], especially under a few rare event trial settings [175]‡.

Examples of meta-analysis in orphan drug evaluation

Dealing with heterogeneity – (Bronchitol 2012)

Usually, in an orphan drug evaluation procedure, no more than two randomized trials are available [100]. This was the case with the evaluation of Bronchitol® (mannitol) for cystic fibrosis [177]. The pivotal study (DPM-CF-301) resulted in a significant effect of absolute change from baseline in Forced Expiratory $Volume_1$ over 26 weeks of 54.17 ml (95% CI: 24.73,83.60), while the second pivotal study (DPM-CF-302) did not achieve statistical significance (54.14 ml (95% CI: -1.97,110.26)) when compared to a sub-therapeutic dose of mannitol. Concerns were raised upon the data quality due to high rates of drop-outs in trial DPM-CF-301. A post-hoc baseline corrected analysis of trial DPM-CF-302 resulted

in significant results (71.10 ml (95% CI: 19.11,123.09). The second study showed greater uncertainty around the treatment effect [177]. Even though statistical heterogeneity was not observed, Bronchitol indicates that methodological heterogeneity can be introduced between two individual studies, in spite of their design being identical. In such cases, a pooled analysis has to be performed with caution [108].

Individual patient data meta-analysis – (Wakix 2015)

Within a prospectively planned drug evaluation, IPD MA can facilitate a thorough investigation of possible sources of heterogeneity through full regression and subgroup analysis in the light of only a few trials. According to the Wakix® EMA report for the treatment of narcolepsy, inferences were different between the two individual pivotal trials based on significance testing [112]. It was suggested that this occurred due to the different maximum dosage levels of the active treatment. A post-hoc individual patient data analysis was applied and showed efficacy in all dosage levels of the active treatment. Wakix report does not indicate whether the between-study variance was accounted for in the pooled analysis, but highlights the well known advantages of exploring heterogeneity via IPD. Such an IPD analysis could have been performed in all aforementioned examples but becomes more relevant in cases of inconsistent inferences [108].

Meta-analysis of adverse events – (Mozobil 2009, Darzalex 2016)

Crude analysis of safety outcomes by using aggregated tables is a common procedure for a drug evaluation [111, 113]. A formal meta-analysis is usually not performed. The Mozobil® EMA report presents only crude safety tables of the two Phase III studies AMD3100-3101 and AMD3100-3102. In both studies 600 patients enrolled and were equally allocated between plerixafor and placebo. For example, during period 1, diarrhoea, a common adverse event, and deep venous thrombosis, a serious adverse event, were both pooled and reported as 37.6% vs. 16.6% and 1.34% vs. 0.03% for plerixafor vs. placebo respectively. The same report compares the average percentages of all serious adverse events (as reporting at least 1 serious adverse event during period 1) and concludes that they are comparable between the two treatment arms over the pooled studies (4% plerixafor vs. 5.8% placebo). The above change in the treatment safety between individual adverse events and the simply pooled safety, may

be the result of Simpson's paradox [149].

Such a paradox could be produced under the naive pooling studies 3101 and 3102 or the naive pooling of different adverse events, especially when adverse events are not very rare [149]. This practice probably dominated since the combination of a few small studies with zero events results in an unstable frequentist MA [43, 149, 150]. In this case, the Bayesian framework offers a flexible alternative for synthesizing rare adverse events across a few small trials, since it requires no continuity correction when zero-event studies are included and no normal approximation of the data distribution as opposed to the frequentist MA ([172, 173]‡ and [178]).

Network meta-analysis – (Torisel 2007)

In most rare diseases multiple treatments are not available. A more relevant application of NMA within an orphan drug evaluation may be the comparison of accumulated evidence among alternative dosages or drug combinations. More specifically, the Torisel® (temsirolimus) EMA report presents two randomized studies that compare different dosages of the main intervention (Temsirolimus "25mg", "75mg", "250mg", "Interferon-alpha" and "Temsirolimus 15mg/Interferon-alpha") for the treatment of renal cell carcinoma [179]. Treatment "Temsirolimus 25mg" is reported in both studies. In the same example, even though a direct comparison between "Interferon-alpha" and "Temsirolimus 75mg" does not exist, an indirect effect estimate could be calculated via their common comparator "Temsirolimus 25mg". However, this indirect estimate is produced solely by two trials, and hence there might not be sufficient power to detect a treatment effect, as compared to a direct estimate for the same comparison [139]. This practice has been described in the context of a non-inferiority trial, where the standard treatment has been compared both with placebo and the new treatment, but the latter has been compared only with the standard treatment [180]‡.

Discussion

We have identified 31 articles that describe methods of combining a series of trials for rare diseases within the Bayesian framework. We focused on synthesis methods for available studies, excluding the use of a meta-analysis for the design of a future study. As expected,

very few studies dealt directly with small number of trials and/or small sample sizes. Since research on meta-analysis methods that are specifically tailored to small populations were only recently explored ([13, 147])‡, our goal was to illustrate potential directions and research areas of interest for small populations and to comment on their applicability in a rare disease drug evaluation.

In the light of a limited number of studies, conclusions based on a series of different yet relevant trials (i.e. an implication of RE MA) have been regarded as particularly suboptimal within a frequentist context [11]. Nonetheless, Bayesian methods are not frequently applied and this seems to be the case for both large-scale and rare diseases. Especially for rare diseases, the fixed-effects inverse-variance and the Mantel-Haenszel methods are commonly applied instead [181, 182]. The assumption of fixed-effects for estimation contradicts the usual heterogeneous nature of rare conditions that do work. Indeed, the proper estimation of heterogeneity is an unresolved issue for classical MA that consists of only two trials ([11]‡ and [13]) and has led some to advocate against any formal synthesis ([26]‡ and [183, 184]). This inability of frequentist methods to account for heterogeneity in such settings led authors to advocate in favour of Bayesian approaches in MA instead [13, 147]‡. Bayesian approaches, through Markov chain Monte Carlo (MCMC) computational methods, have become even more popular by eliminating compromises when modelling (i.e. assuming a normal distribution for data that are clearly non-normal), while they are availabe in streamlined statistical programs [185, 186].

A Bayesian MA utilizes historical information through prior distributions on model parameters. Such informative priors become necessary, particularly when evidence is sparse, a setting that leads to the improper estimation of heterogeneity. The utilization of priors that cover a plausible range for τ ([131] and [13, 147])‡ or published predictive distributions as priors in a MA could be employed as a possible solution ([151]) and [152, 153, 154, 155, 156]‡). Nonetheless, applying the latter prior distributions to rare diseases might be risky, as they may not truly represent the degree of heterogeneity of such diseases. These priors are produced mostly through Cochrane reviews which have been recently shown to provide systematically different results from the non-Cochrane reviews in terms of magnitude and significance [187]. We suggest either an investigation via simulation on -already proposed-

predictive distributions in order to evaluate their performance in the light of rare diseases or an approximate refinement (i.e. through variance downweighting), in order to improve their performance in rare diseases. In addition, we suggest constructing similar predictive distributions that are based on the disease's prevalence, which is expected to impact the degree of heterogeneity.

To our knowledge, NMA has not been incorporated yet in a rare disease evaluation of the EMA. However, based on the method's ability to synthesize direct and indirect evidence of several competing treatments in a single analysis, NMA could play a role in the drug development process. Even though it was outside from this review scope, for example, prior to initiating a new trial, a formal synthesis of all treatments that are available for a specific disease may be performed, so as to inform the design of new trials [188], namely, to decide which pairwise comparison(s) needs (further) investigation and to inform the design characteristics of the required trials (e.g. number of patient, clinically worthy treatment effect) [189]. Recently, this concept was expanded to introduce the notion of living cumulative NMA, which aims to provide a constantly updated meta-analysis of any available treatments in a therapeutic area [190, 191]. Applications of NMA in rare diseases have already emerged in the literature i.e. multiple myeloma [136] and cystic fibrosis [137].

Before applying complex methods, such as MMA or NMA, practitioners should consider that the risk of decreased efficiency is higher than in a PMA due to the additional parameters and the underlying assumptions ([26][‡] and [10, 141]). To evaluate the applicability of the additional assumptions that characterize such models in rare diseases, simulation studies are needed in scenarios of sparsely connected networks and limited available data. Similar shortcomings appear for models that synthesize different study designs and add an extra level of model complexity; for example, the synthesis of randomized and non-randomized evidence, a topic that has been discussed extensively [192].

Based on this review, we did not identify any IPD MA methods that are tailored for small populations. In general, when the size and number of trials are inadequate for conducting subgroup analyses or exploring interaction effects, then inferences on the overall treatment effect of an AD MA and an IPD MA are not expected to differ significantly [160, 161, 193]. IPD in MA and NMA offers several advantages, namely they; (1) provide the opportunity

to explore differences between subgroups more efficiently, (2) facilitate a decent evaluation of heterogeneity sources and (3) ensure a credible handling of missing outcome data [94, 161]. To our knowledge, IPD MA in the context of rare diseases has never been evaluated in a single simulation study. In such a study, the one- and two-stage principles would be compared over a sensible set of simulation scenarios, that are tailored for characteristics of rare diseases. Such a study could offer insight on the importance of a prospective IPD MA in a orphan drug development process as opposed to a retrospective IPD MA (i.e. IPD availability is attainable in the former). In case of no access to IPD for all studies, we could increase efficiency in rare diseases by properly combining IPD, AD and relevant non-randomized study data. To increase IPD availability, commercial companies should further initiate a data-sharing mechanism. Such initiatives are either in discussion or already active [194, 195].

Finally, we should acknowledge a number of limitations in this article. This scoping review was not meant to be extensive, nonetheless, we evaluated articles that came from 5 search engines as this can provide insights on future search strategies for the interested readers (Supplementary material 1). Our search was built on an 9-year time frame supplemented by an ad-hoc monitoring of increased methodological interest journals, which we believe provided us with enough articles and enabled us to formulate directions. Only a few articles dealt directly with meta-analytical approaches in rare disease settings [13, 147]‡. Nonetheless, the rest of the literature addressed issues of interest for rare diseases in an indirect manner (i.e. rare events, few number of studies, estimation of heterogeneity). Throughout the manuscript, the term "small populations" was used to refer to rare diseases, however, small population conditions can appear in subgroups of common diseases. Thus, part of our findings could apply to this setting as well.

In this scoping review we focused primarily on Bayesian methods for meta-analysis and briefly discussed frequentist methods, mainly due to the reported deficiencies of frequentist MA in rare diseases and appealing characteristics of a Bayesian MA ([11, 147]‡ and [150]). Even though most approved orphan drugs show that the conduct of reasonably sized trials is possible during an orphan drug development [100], this characteristic usually represents only the most prevalent among rare diseases. In very and ultra rare diseases, no more than two small trials become available. Our objective was to identify possibly relevant directions

for Bayesian synthesis methods in orphan drug evaluations and we trust this has been accomplished to a large extent despite the above limitations.

Conclusions

To our knowledge, this is the first review discussing aspects of Bayesian meta-analysis in rare diseases. Bayesian meta-analysis methods may overcome efficiently methodological difficulties inherent to evidence scarcity. Nevertheless, available Bayesian meta-analysis methods tailored to small populations are currently not common. Marketing authorization of orphan drugs could benefit from the Bayesian methods in the context of series of small trials but more methodological developments are needed for the application of Bayesian meta-analysis in an orphan drug evaluation.

Acknowledgements

The authors would like to thank Maria Fotsala for extensive textual and language editing.

Supplementary material can be found at `figshare.com/s/63cab08581021a87417b - 10.6084/m9.figshare.11977794` or/and at the online manuscript.

Chapter 3

Data generating models of dichotomous outcomes: Heterogeneity in simulation studies for a random-effects meta-analysis

K Pateras
S Nikolakopoulos
KCB Roes

Statistics in Medicine. 2018 ; 37(7): 1115-1124.
https://doi.org/10.1002/sim.7569

Chapter 3

Abstract

Simulation studies to evaluate performance of statistical methods require a well specified Data Generating Model. Details of these models are essential to interpret the results and arrive at proper conclusions. A case in point is random-effects meta-analysis of dichotomous outcomes. We reviewed a number of simulation studies that evaluated approximate normal models for meta-analysis of dichotomous outcomes and we assessed the data generating models that were used to generate events for a series of (heterogeneous) trials. We demonstrate that the performance of the statistical methods, as assessed by simulation, differs between these three alternative data generating models, with larger differences apparent in the small population setting. Our findings are relevant to multilevel binomial models in general.

3.1 Introduction

Increasingly, simulation studies are used to assess properties of statistical methods in more complex settings. In addition to the statistical models used, the actual operational model and methods with which data are generated can impact the results. The Data Generating Model (DGM) [196] is essential to interpret the results, to arrive at proper conclusions and to compare between different simulation studies. More often than not, in our simulation work we returned to the details of this DGM to understand results and sometimes correct to better fit the statistical model and realistic scenarios. A case in point is the random-effects meta-analysis of dichotomous outcomes, towards which recently several simulation based research papers addressed different questions, particularly for a few or small trials [11, 16, 197].

The standard model for random-effects meta-analysis assumes approximately normal effect estimates $Y_i \sim N(\theta_i, s_i^2)$, for trial $i = 1, ...k$ for the study-specific effects θ_i and a normal-normal hierarchical model around the study effects $\theta_i \sim N(\theta, \tau^2)$, where s_i^2 are the study-specific within-study variances and τ^2 is the between-study variance. In the case of dichotomous outcomes we can model the study-specific effects θ_i as the $log(OR) = logit(pT) - logit(pC)$, where pT is the experimental treatment arm event rate and pC the control arm event rate. Evidently, the normal approximation to the binomial distribution breaks down in the case of small samples or small number of events and this can have consequences for the DGM and its utilization in simulation studies.

Simulations of (individual) trial data in this setting, particularly for small trials, would typically generate numbers of events per trial arm according to binomial distributions, given pT and pC. However, the additional between-study variability implied by the (approximate) normal-normal model in this case should now be incorporated in modelling pT and pC, which a priori can be done in different ways. We reviewed a number of simulation studies that used the normal-normal model for dichotomous outcomes [11, 16, 27, 74, 144, 147, 197, 198, 199, 200] and assessed the DGMs used to produce event rates (pT, pC) and generate events for a series of (heterogeneous) trials. In section 3.2 we present and discuss the DGMs. In section 3.3 we perform a comparison of the DGMs under three widely applied meta-analytical models via a simulation study. The manuscript concludes with a discussion in section 3.4.

3.2 Data generating models

In the literature, so far, at least three alternative DGMs were utilized for generating individual trial data. The first makes the assumption of homogeneity in the control arm and places all the between-study variance in the event rate pT of the treatment arm [74, 144]; we refer to this as "$pCFixed$". The second is based on the assumption of a fixed average trial risk ($p_i0 = (p_{iT} + p_{iC})/2$), with which we calculate the event probability in each arm, based on a simulated overall treatment effect [16, 200]; we refer to this as "$pAverage$". The third is based on the incorporation of the between-study variance in both treatment arms via the use of logits [11, 198]; we refer to this as "$pRandom$". The steps to generate events for each DGM are presented below.

Algorithm 1 - Data Generating Model $pCfixed$

1: Set θ,τ, a range for p_{iC} and a range for m_i, $i = 1, ..., k$ and $j = (C)ontrol, (T)reatment$.

2: $m_i \sim Uniform(m_{lo}, m_{up})$ - Generate study-arm sample sizes.

3: n_{ij} = m_i - Set equal study-arm allocation ratios.

4: $\theta_i \sim Normal(\theta, \tau)$ - Generate study-specific treatment effects.

5: $p_{iC} \sim Uniform(\alpha, \beta)$ - Generate a study-specific control event probability.

6: $p_{iT} = p_{iC} \cdot exp(\theta_i)/(1 - p_{iC} + p_{iC} \cdot exp(\theta_i))$ - Compute the study-specific treatment event probability.

7: $r_{ij} \sim Binomial(p_{ij}, n_{ij})$ for $j = C$ and T - Generate study events.

Algorithm 2 - Data Generating Model $pAverage$

1: Set θ,τ, a range for p_i0 and a range for m_i, $i = 1, ..., k$ and $j = (C)ontrol, (T)reatment$.

2: $m_i \sim Uniform(m_{lo}, m_{up})$ - Generate study-arm sample sizes.

3: $n_{ij} = m_i$ - Set equal study-arm allocation ratios.

4: $\theta_i \sim Normal(\theta, \tau)$ - Generate study-specific treatment effects.

5: $p_i0 \sim Uniform(\alpha, \beta)$ - Generate a study-specific average event probability.

6: $p_i0 = \sum_{j=1}^{2} p_{ij}/2$

7: $\theta_i = log\Big(\frac{(p_{iC}) \cdot (1 - p_{iT})}{(p_{iT}) \cdot (1 - p_{iC})}\Big)$.

8: Solving (6) and (7) we acquire p_{ij}.

9: $r_{ij} \sim Binomial(p_{ij}, n_{ij})$ for $j = C$ and T - Generate study events.

Algorithm 3 -Data Generating Model $pRandom$

1: Set θ, τ, $p_{iC,Init}$ and a range for m_i, $i = 1, ..., k$ and $j = (C)ontrol, (T)reatment$.

2: $p_{iT,Init} = p_{iC,Init} \cdot exp(\theta)/(1 - p_{iC,Init} + p_{iC,Init} \cdot exp(\theta))$ - Compute the initial study-specific treatment event probability.

3: $m_i \sim Uniform(m_{lo}, m_{up})$ - Generate study-arm sample sizes.

4: $n_{ij} = m_i$ - Set equal study-arm allocation ratios.

5: $\mu_{ij} = log(p_{ij,Init}/1 - p_{ij,Init})$ - Compute mean logits given initial fixed event rates $p_{ij,Init}$.

6: $logit_{ij} \sim Normal(\mu_{ij}, \tau/\sqrt{\mathbf{2}})$ - Generate study-specific control and treatment logits.

7: $p_{ij} = \frac{1}{1 + e^{-logit_{ij}}}$ - Back-calculate the event rates for each trial arm.

8: $r_{ij} \sim Binomial(p_{ij}, n_{ij})$ for $j = C$ and T - Generate study events.

Note that for two of the DGMs discussed, the use of Uniform distributions is utilized ($pCFixed$ and $pAverage$). This is done in order to replicate their use in the literature [16, 74, 200] . This adds an additional source of variability, not specifically modelled by the normal-normal hierarchical model. We keep using the term ”fixed” and ”homogeneous” for these DGMs, even if the probability of events is not kept fixed across studies. We retain the term ”heterogeneous” for referring to heterogeneity resulting from the variance parameter of the random-effects model.

As Figure 3.1 demonstrates, the primary difference among the three presented DGMs lies in the joint distribution of the two model event rate parameters, as used in generating data. Homogeneity of the control group event rates ($pCFixed$) has been discussed previously [201] and can be observed in the densities of Figure 3.1. The study-specific control event rates are homogeneous - coming from a $Uniform(0.1, 0.3)$- , while the study-specific treatment

event rates are heterogeneous. The $pAverage$ approach makes an intuitively restrictive assumption since it constrains the simulated values of the control and treatment arm around an average true risk rate. The $pRandom$ approach places the between-study variability in both treatment arms without imposing additional constraints. In this DGM, it is common to assume equal between-study standard-deviation ($\tau/\sqrt{\mathbf{2}}$) in both arms, an assumption which might not always hold in practice, but can be relaxed. For example, the standard of care - control treatment- might be less variable between studies in comparison to the experimental treatment; or more variable if the standard of care differs between regions or countries, a flexibility that is not straightforward to implement in the other two DGMs discussed here.

Indeed, in the $pCFixed$ DGM, the probability of events in the two arms is not fixed, but rather randomly generated via a joint distribution at the study-parameter level, where the control group rate is considered to be independent from the effect size. The $pRandom$ DGM is largely the same, except that the range of pC is not restricted, and its distribution is skewed within its range. Naturally, after incorporating smaller heterogeneity in the control group, $pCFixed$ can be considered a special case of $pRandom$ if the two parameters of the Uniform distribution generating event rates in $pCFixed$ are equal (Algorithm 3 - Step 6).

An important difference between the presented DGMs arises from their ability to accommodate ranges of event rates. The $pAverage$ directly defines the average event rate ($p0$), the $pCFixed$ directly defines the control group rate (pC), whereas the $pRandom$ does not allow a direct impact on event rates. These fundamental characteristics of the three DGMs render their fair comparison through simulation less trivial. For the specific scenario studied here, where probabilities of events on the control groups are smaller than 50%, whenever the average effect size is positive, the control group event rate for $pAverage$ is, on average, smaller than the competing DGM's simulations. This implies smaller numbers of events in the two arms. The total number of events is related to power. Therefore differences in empirical power of the $pAverage$ method may appear partially due to this difference in the average rate. Nonetheless, the constraints of the $pAverage$ DGM inherently restrict the DGM from jointly exploring very low event rates (Figure 3.1), which minimizes the event rate's impact on power. Thus, the $pAverage$ DGM makes (empirical) very large effect sizes less probable. This restriction becomes problematic particularly for studies in small populations where we

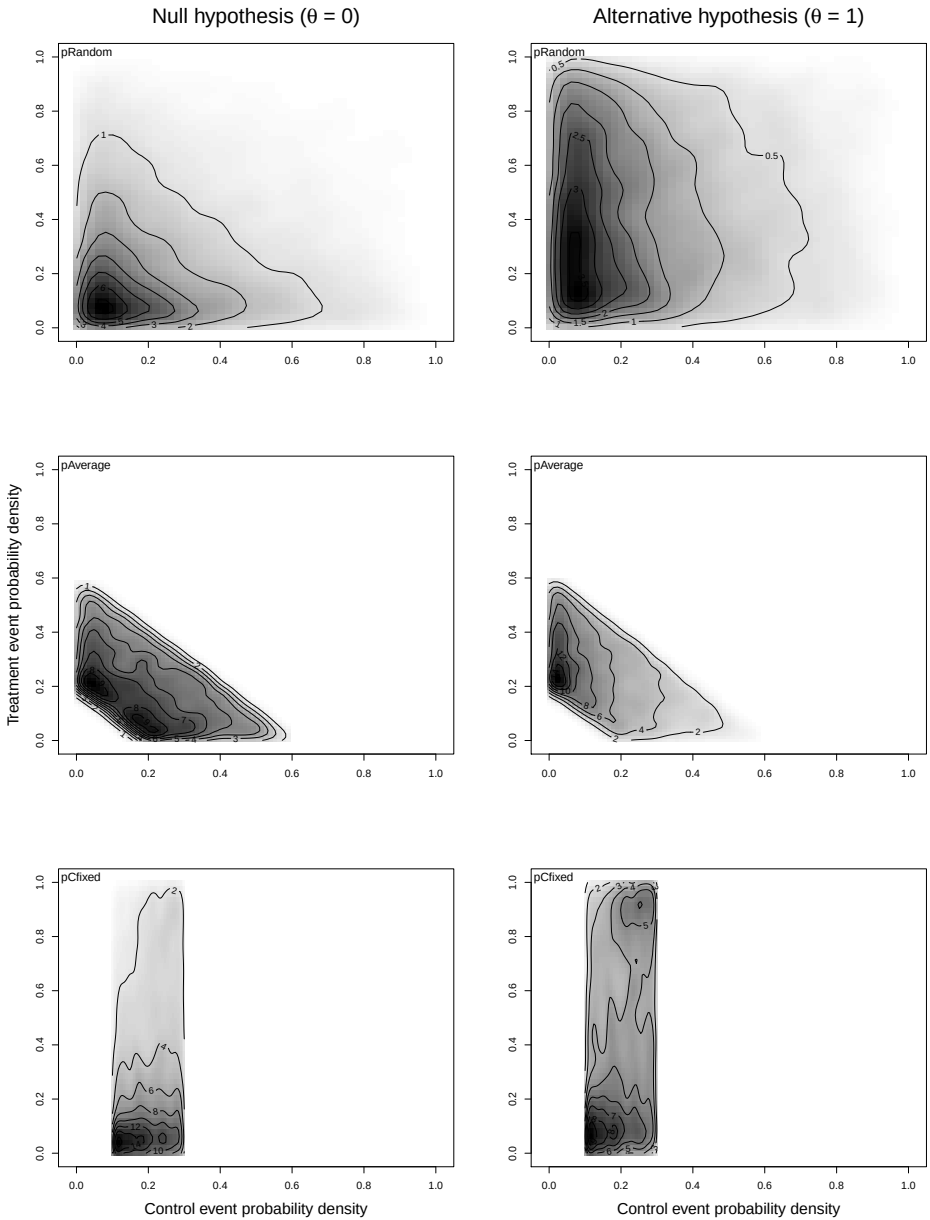

Figure 3.1: Empirical numerically estimated joint event probability densities for the control and treatment arm of the three Data Generating Models under the null and alternative hypothesis with substantial between-study standard-deviation ($\tau = 2$), small sample size ($n_{ij} = m_i \sim Uniform(20, 30)$, $i = 1, 2\ ;\ j = Control, Treatment$) and an (average) event rate, either as a fixed value of 0.20, or as a mean of 0.20 of a $Uniform(0.1, 0.3)$ distribution.

usually seek to observe very large effect sizes.

Ideally, results and conclusions of simulation studies are expected to hold for the statistical model specified, and not to depend on the characteristics of the utilized DGM. Since they all generate the same true overall treatment effect, when we use statistical methods in the setting of many large trials and relatively frequent events, we may expect similar results under different DGMs. However, in the case of a small number of trials with small sample sizes, the normal approximation of the logOR might be insufficient and more sensitive to the choice of DGM. Thus, possible differences between the observed performance of methods may be enhanced.

Evidence from a few small trials would often become available or would be sufficiently similar to be synthesized, during a drug development and evaluation in rare diseases [182, 202, 203]. Until recently, the evaluation of meta-analytical methods in this rare disease context was not common. However, more attention has been drawn to this topic, especially since the initiation of three European projects, focused on characteristics of statistical methodologies in small populations **(ASTERIX, IDeAl and InSPiRe)**. A number of articles have now been published that evaluate methods for a meta-analysis of a few and even two small trials [11, 147, 197].

3.3 Simulation study and results

To compare the implications of the three DGMs for both a meta-analysis of two large trials ($n_{ij} = m_i \sim Uniform(230, 240), i = 1, 2\ ;\ j = (C)ontrol, (T)reatment$) and a meta-analysis of two small trials ($n_{ij} = m_i \sim Uniform(20, 30), i = 1, 2\ ;\ j = C, T$), we follow Gonnermann et al [11] to evaluate the statistical properties of three meta-analysis methods; (1) a fixed-effect meta-analysis (FE), (2) a random-effects meta-analysis with the application of DerSimonian and Laird heterogeneity estimator [204] (DL) and (3) a random-effects meta-analysis with the Hartung and Knapp correction [198] (HK) for a meta-analysis of two trials. The (average) event rate is assumed to be 0.20. This is, however, interpreted and implemented differently between the DGMs used, i.e., either as a fixed value of 0.20 ($pRandom$), or as a mean of 0.20 of $Uniform(0.1, 0.3)$ on either pC ($pCFixed$) or $p0$ ($pAverage$). We also apply a continuity correction (0.5) in all cells of a trial with zero cells. We assume equal allocation ratios within each trial. We present results under the null and the alternative hypothesis with varied

levels of true between-study standard-deviation $\tau \in \{0.001, 0.5, 1, 2\}$, which corresponds to relative heterogeneity of $I^2 \approx \{0.01\%, 47\%, 63\%, 75\%\}$ for a small trial meta-analysis and $I^2 \approx \{0.05\%, 74\%, 84\%, 90\%\}$ for a large trial meta-analysis.

Table 3.1: Empirical type I error and empirical power based on 10^6 simulations. PR: $pRandom$, PA: $pAverage$, PCF: $pCFixed$, FE: Fixed-effect approach, HK, Hartung and Knapp approach, DL: DerSimonian Laird approach, θ: overall treatment effect (log odds ratio), τ: between-study standard-deviation, Small sample size: $n_{ij} = m_i \sim Uniform(20, 30)$, Large sample size: $n_{ij} = m_i \sim Uniform(230, 240)$, $i = 1, 2$; $j = Control, Treatment$

			Empirical type I error ($\theta = 0$)			Empirical power ($\theta = 1$)		
Method	S.size	τ	PR	PA	PCF	PR	PA	PCF
DL	Small	0.001	0.029	0.028	0.028	0.482	0.340	0.465
		0.5	0.062	0.058	0.060	0.417	0.311	0.420
		1	0.133	0.118	0.125	0.341	0.280	0.357
		2	0.224	0.210	0.205	0.297	0.278	0.298
	Large	0.001	0.037	0.038	0.038	0.998	0.986	0.997
		0.5	0.207	0.201	0.202	0.759	0.736	0.757
		1	0.267	0.264	0.263	0.487	0.476	0.489
		2	0.288	0.286	0.284	0.350	0.345	0.351
HK	Small	0.001	0.047	0.048	0.048	0.132	0.114	0.129
		0.5	0.047	0.049	0.049	0.106	0.098	0.107
		1	0.050	0.050	0.051	0.080	0.081	0.083
		2	0.056	0.059	0.059	0.065	0.078	0.066
	Large	0.001	0.052	0.052	0.052	0.400	0.339	0.394
		0.5	0.050	0.050	0.050	0.162	0.154	0.163
		1	0.050	0.049	0.049	0.092	0.089	0.092
		2	0.055	0.050	0.049	0.065	0.061	0.061
FE	Small	0.001	0.033	0.032	0.032	0.572	0.388	0.547
		0.5	0.075	0.070	0.075	0.532	0.369	0.539
		1	0.186	0.151	0.191	0.486	0.345	0.537
		2	0.373	0.270	0.391	0.480	0.346	0.553
	Large	0.001	0.049	0.049	0.049	1.000	0.999	1.000
		0.5	0.397	0.377	0.390	0.968	0.947	0.967
		1	0.627	0.570	0.632	0.857	0.800	0.883
		2	0.767	0.641	0.803	0.816	0.696	0.872

Table 3.1 presents the differences between the empirical power curves when a treatment effect is present ($\theta = 1$) and under the null hypothesis ($\theta = 0$) for each DGM and each considered meta-analytical method. The $pAverage$ DGM produces data that result in lower power than the other two DGMs, especially for the FE and DL approach. This can be explained by the constraints that are induced on the event probabilities (Figure 3.1), which are in turn influenced by the specific choice of $\alpha = 0.1$ and $\beta = 0.3$ for the Uniform distributions. In addition, regarding high levels of true heterogeneity, the $pCFixed$ tends to increase the empirical power of the FE approach. This could be expected, as when $\tau^2 > 0$ and heterogeneity is only applied to the treatment group event rates, larger effect sizes are produced compared to the other two DGMs. In terms of type I error, empirical values seem to be heavily dependent on the DGM when the FE approach is assessed. Evaluation of the Hartung and Knapp approach is less affected by the DGM, with small deviations in empirical power, and mostly for the $pAverage$ DGM. A graphical representation of the empirical power curves can be found in Figures A1 and A2.

Figure 3.2 summarizes the performance of the three DGMs in terms of coverage of the 95% confidence intervals for small sample sizes, $m_i \sim Uniform(20, 30)$. Under heterogeneous conditions ($\tau = 1$), especially for the FE and DL approaches, the $pRandom$ demonstrates lower coverage than the $pCFixed$ and $pAverage$, across the considered levels of overall treatment effect. Regarding large sample sizes ($m_i \sim Uniform(230, 240)$), in terms of coverage of the 95% confidence intervals, the three DGMs show similar behaviour for homogeneous conditions (Figure A3). On the contrary, for heterogeneous conditions the $pAverage$ DGM starts to favour the FE and DL approaches when $\theta \geq 2$, bringing the three methods 95% coverage relatively closer than $pCFixed$ and $pRandom$.

3.4 Discussion

The choice of a DGM used in simulation studies is important and has to be consistent with the assumed statistical model under realistic assumptions related to the issue in question. Our simulations show that statistical methods perform differently across DGMs that were used to investigate properties of random-effects meta-analyses. Our simulation is not extensive and does not cover effects in other settings. Nonetheless, we noticed that the divergent behaviour

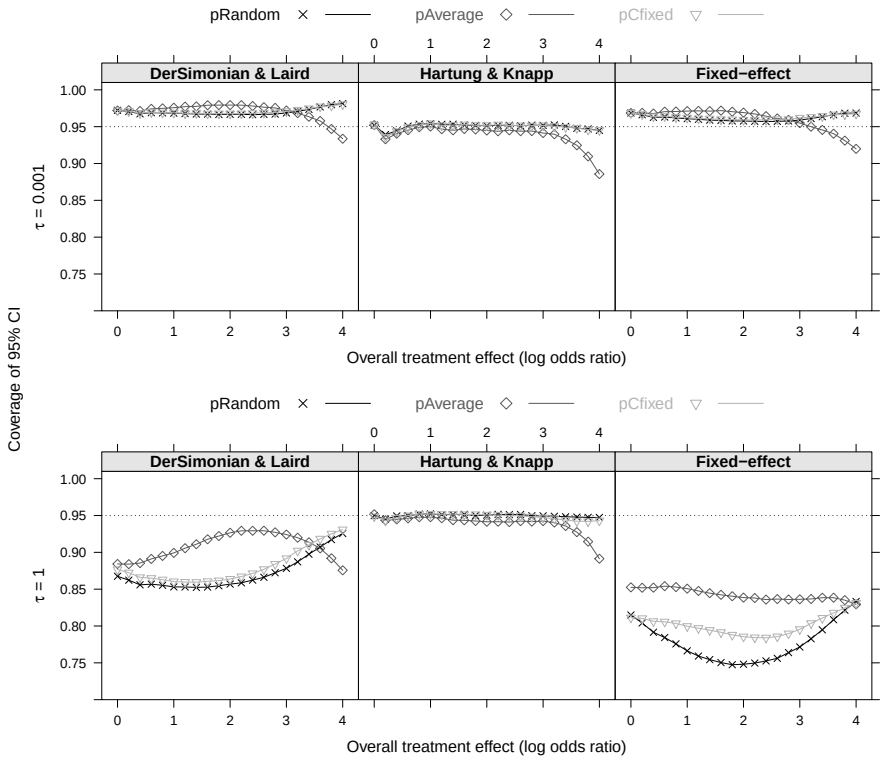

Figure 3.2: Impact of data generating mechanism in a meta-analysis of two small studies ($n_{ij} = m_i \sim U(20, 30)$, $i = 1, 2$; $j = Control, Treatment$) on coverage of the 95% confidence intervals. τ: between-study standard-deviation.

of DGMs is preserved when synthesizing many small trials, but is reduced when synthesizing many large trials. In contrast to large study meta-analyses simulation studies, the choice of a DGM can impact the conclusions of small study meta-analyses simulation studies to a greater extent. The findings actually extend beyond the presented small population context and hold more generally for multilevel binomial data settings.

The elaboration on the DGM articulates one of the crucial conceptual difficulties of the random-effects model for meta-analysis. In all three random-effects DGM formulations and assessments of type I errors, in the presence of heterogeneity, there is also heterogeneity under the null hypothesis. Although all three DGMs are designed to produce the same true overall effect, the properties of the modelled joint empirical distribution of the control and treatment

event rates can differ dramatically.

As a consequence, simulation studies that use different DGMs for essentially the same overall statistical model have the potential to result in different conclusions regarding performance of the statistical methods investigated. For this reason, methodological reviews for meta-analysis [205, 206] have to report in detail the DGM of each study they include and potential consequences of the choice of DGM. If flexible assumptions on the event probability are needed, the use of $pRandom$ DGM might be recommended. We believe that not enough emphasis is placed on the proper choice nor on the sufficient reporting of DGMs in both individual simulation studies and methodological reviews.

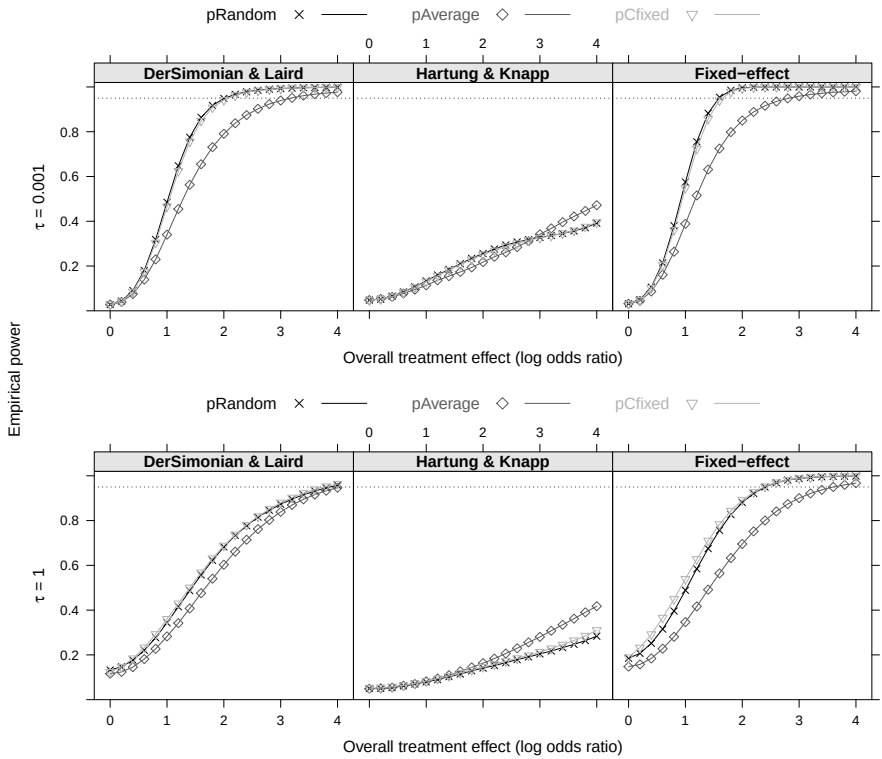

Figure A1: Impact of data generating mechanism in a meta-analysis of two small studies ($n_{ij} = m_i \sim Uniform(20, 30)$, $i = 1, 2$; $j = Control, Treatment$) on empirical power. τ: between-study standard-deviation.

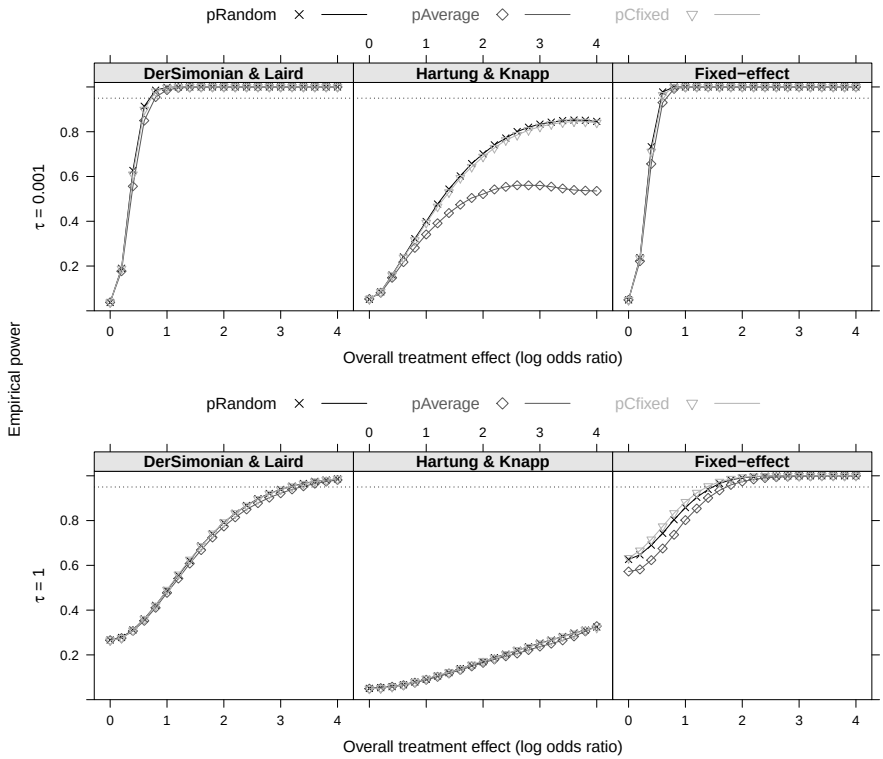

Figure A2: Impact of data generating mechanism in a meta-analysis of two large studies ($n_{ij} = m_i \sim Uniform(230, 240)$, $i = 1, 2$; $j = Control, Treatment$) on empirical power. τ: between-study standard-deviation.

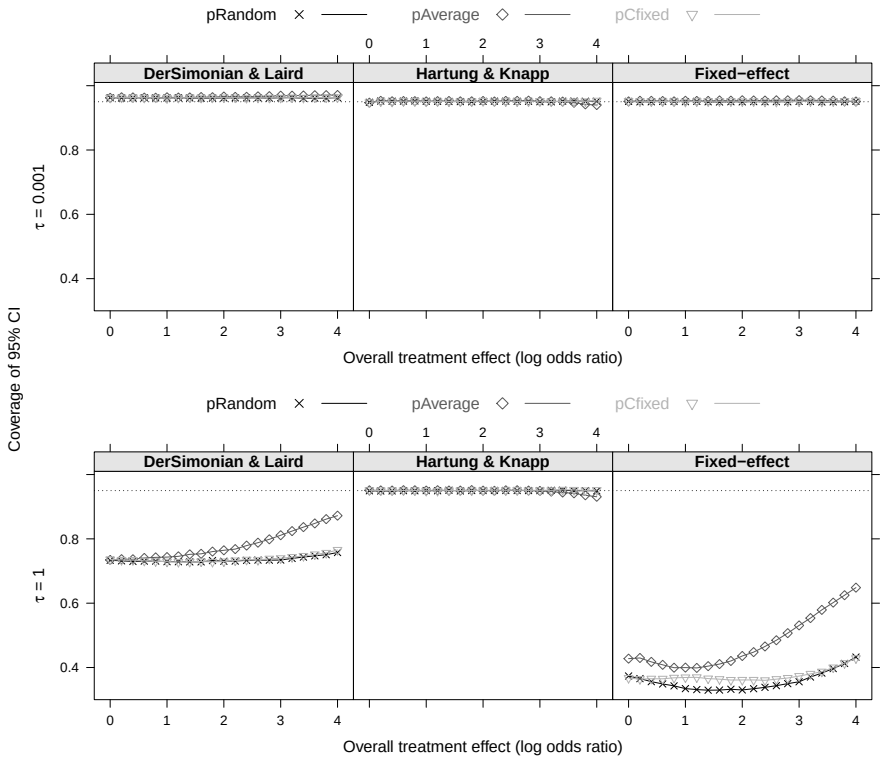

Figure A3: Impact of data generating mechanism in a meta-analysis of two large studies ($n_{ij} = m_i \sim Uniform(230, 240)$, $i = 1, 2$; $j = Control, Treatment$) on coverage of the 95% confidence intervals. τ: between-study standard-deviation.

Chapter 4

Interval estimation of the overall treatment effect in a meta-analysis of a few small studies with zero events

K Pateras
S Nikolakopoulos
D Mavridis
KCB Roes

Contemporary Clinical Trials Communications. 2018 ; (9): 98–107.
https://doi.org/10.1016/j.conctc.2017.11.012

Chapter 4

Abstract

When a meta-analysis consists of a few small trials that report zero events, accounting for heterogeneity in the (interval) estimation of the overall effect is challenging. Typically, we predefine meta-analytical methods to be employed. In practice, data poses restrictions that lead to deviations from the pre-planned analysis, such as the presence of zero events in at least one study arm. We aim to explore heterogeneity estimators behaviour in estimating the overall effect across different levels of sparsity of events. We performed a simulation study that consists of two evaluations. We considered an overall comparison of estimators unconditional on the number of observed zero cells and an additional one by conditioning on the number of observed zero cells. Estimators that performed modestly robust when (interval) estimating the overall treatment effect across a range of heterogeneity assumptions were the Sidik-Jonkman, Hartung-Makambi and improved Paul-Mandel. The relative performance of estimators did not materially differ between making a predefined or data-driven choice. Our investigations confirmed that heterogeneity in such settings cannot be estimated reliably. Estimators whose performance depends strongly on the presence of heterogeneity should be avoided. The choice of estimator does not need to depend on whether or not zero cells are observed.

4.1 Introduction

Meta-analyses (MAs) techniques are commonly employed in order to obtain a more precise and more general effect estimate of a treatment. Heterogeneity (τ) of treatment effects measured in multiple Randomized Controlled Trials (RCTs) is a crucial part of the estimation [26].

In MAs of RCTs, methodological challenges arise when the disease under examination is rare and only a few small RCTs are available [99, 100]. This is mostly due to the large sample assumptions on which most MA methods are based. In the case of rare diseases with binary endpoints, zero cells are more likely to be observed in at least one of the treatment arms of at least one contributing trial [182, 207, 208]. Zero cells in MAs pose challenges as they induce bias in both the estimation of the overall effect and the between-study variance (heterogeneity) [43, 149, 209, 210, 211, 212, 213, 214].

When conducting a MA, the estimation method might be adjusted conditionally on observing zero cells. Corrections are typically introduced by adding a number to the zero cells observed; furthermore, the heterogeneity estimator could change. The latter choice is by itself a challenging task, given the large pool of options [92, 115, 159, 204, 215, 216, 217, 218, 219, 220].

Especially for dealing with a MA of a few RCTs, there is no straightforward answer to which estimator would be robust across several heterogeneity assumptions [159]. Most estimators face difficulties in case of a limited number of trials; they induce bias in the estimation of τ [221, 222] and may result in inappropriate interval estimation of the treatment effect. However, not much is known regarding their behaviour in the presence of zero cells and small populations.

The primary objective of this work is to assess the robustness of heterogeneity estimators in the (interval) estimation of treatment effect across ranges of sparsity of events and assumed heterogeneity. The starting point is the acknowledged poor estimation of heterogeneity in this setting. We evaluate the estimators in case they are predefined (unconditional), as well as when they are chosen depending on the observed zero cells in contributing trials (conditional on the observed data, in short: conditional).

The paper is organized as follows. First we describe the standard random-effects (RE) model and introduce the heterogeneity estimators briefly. Subsequently, we present two motivating examples and their analysis. Then we describe the simulation study and evaluate the two distinct approaches. We conclude with recommendations on evidence synthesis for a sparse-events MA in small populations.

4.2 Methods

We consider a set of k trials with binary outcomes that compare an experimental treatment to a control. Patients are randomized between two groups: treatment (T) and control (C).

By Y_i we denote the log odds ratio (logOR) in the i^{th} trial. Following standard theory (e.g. [26]), we assume:

$$Y_i|\theta_i \sim N(\theta_i, \sigma_i^2), i = 1, ..., k \tag{4.1}$$

The study-specific treatment effect estimates are $\hat{\theta}_i = log\Big(\frac{r_{Ti} \cdot (n_{Ci} - r_{Ci})}{r_{Ci} \cdot (n_{Ti} - r_{Ti})}\Big)$, while their variances are $s_i^2 = \frac{1}{r_{Ti}} + \frac{1}{n_{Ti} - r_{Ti}} + \frac{1}{r_{Ci}} + \frac{1}{n_{Ci} - r_{Ci}}$, where r_i and n_i denote the number of responders and the total number of subjects in each trial, respectively.

Assuming a fixed-effects (FE) model, θ is common for all studies ($\theta_i = \theta$). Assuming a RE model, the θ_i are considered exchangeable and follow a normal distribution, that is,

$$\theta_i|\theta, \tau^2 \sim N(\theta, \tau^2) \tag{4.2}$$

where θ is the overall effect and τ^2 is the between-study variance. When $\tau^2 = 0$, then the RE model reduces to the FE model. The pooled effect estimate is calculated as a weighted average $\hat{\theta} = \sum_i w_i Y_i / \sum_i w_i$. The inverse variance (IV) weights are then defined as $w_{i,RE} = 1/(s_i^2 + \hat{\tau}^2)$ for the RE model and as $w_{i,FE} = 1/s_i^2$ for the FE model.

Table 4.1: Summary of heterogeneity estimators, including their equation, abbreviation and source. $w_{i,RE} = \frac{1}{(s_i^2 + \tau^2)}$, $w_{i,FE} = \frac{1}{s_i^2}$, $\bar{Y}_{RE/FE} = \frac{\sum_i w_{i,RE/FE} Y_i}{\sum_i w_{i,RE/FE}}$, $Q_{RE/FE} = \sum_i w_{i,RE/FE}(Y_i - \bar{Y}_{RE/FE})^2$, $c_{RE/FE} = \sum_i w_{i,RE/FE} - \frac{\sum_i w_{i,RE/FE}^2}{\sum_i w_{i,RE/FE}}$, $w_i^* = \frac{1}{(v_{i,ipm}^* + \tau^2)}$, $v_{i,ipm}^* = \frac{1}{n_{(T,i)}+1}\left(e^{-Prc_O - \bar{Y} + \tau^2/2} + 2 + e^{Prc_O + \bar{Y} + \tau^2/2}\right) + \frac{1}{n_{(C,i)}+1}\left(e^{-Prc_O} + 2 + e^{Prc_O}\right)$ Pr_{co}: Observed control event rate, $\tau_O^2 = \sum_i (Y_i - \bar{Y_{FE}})/k$

Methods	Equation	Abbreviation	Source
DerSimonian Laird	$\hat{\tau}_{dl}^2 = max\Big(0, (Q_{FE} - (k-1))/c_{FE}\Big)$	dl	[204]
Positive DerSimonian Laird	$\hat{\tau}_{dl_p}^2 = \hat{\tau}_{dl}^2$ if $\hat{\tau}_{dl}^2 > 0$ and $\hat{\tau}_{dl}^2 = 0.01$ if $\hat{\tau}_{dl}^2 <= 0$	dlp	[215]
Two-step Der Simonian Laird	$\hat{\tau}_{dl2}^2 = max\Big(0, Q_{RE} - (w_{i,RE}^2\, s_i^2 - \frac{\sum_i w_{i,RE}^2\, s_i^2}{\sum_i w_{i,RE}})/c_{RE}\Big)$	dl2	[92]
Hedges	$\hat{\tau}_{he}^2 = max\Big(0, \frac{\sum_i (Y_i - \bar{Y}_{FE})^2}{k-1} - \frac{\sum_i s_i^2}{k}\Big)$	he	[220]
Two step Hedges	Similar to dl2 using the Hedges estimator	he2	[92]
Positive Sidik-Jonkman	$\hat{\tau}_{sj}^2 = max\Big(\frac{\sum_i ((Y_i - \bar{Y}_{FE})^2/(r_i+1))}{k-1}, 0.01\Big)$, $r_i = s_i^2/\hat{\tau}_O^2$	sj	[217]
Model error variance - vc	$\hat{\tau}_{mvvc}^2 = \frac{\sum_i ((Y_i - \bar{Y}_{FE})^2//r_i^*+1))}{k-1}$, $r_i^* = s_i^2/\hat{\tau}_{he}^2$	mvvc	[217]
Paul-Mandel	(τ_{pm}^2) , $F(\tau^2) = \sum_i w_{i,RE}[Y_i - Y_w(\tau^2)]^2 - (k-1)$	pm	[216]
Improved Paul-Mandel	(τ_{ipm}^2) , $F(\tau^2) = \sum_i w_{i,RE}^*[Y_i - Y_w(\tau^2)]^2 - (k-1)$	ipm	[115]
Hartung - Makambi	$\hat{\tau}_{hm}^2 = \frac{Q_{FE}^2}{[2(k-1) + Q_{FE}]c_{FE}}$	hm	[218]
Hunter-Schmidt	$\hat{\tau}_{hs}^2 = max(0, (Q_{FE} - k)/\sum_i w_{i,FE})$	hs	[219]
Maximum Likelihood	$\hat{\tau}_{ml}^2 = max(0, \sum_i w_{i,RE}^2\Big((Y_i - \bar{Y}_{ML})^2 - s_i^2\Big)/\sum_i w_{i,RE}^2)$	ml	–
Restricted Maximum likelihood	$\hat{\tau}_{reml}^2 = max(0, \frac{\sum_i w_{i,RE}^2\Big((Y_i - \bar{Y}_{ML})^2 - s_i^2\Big)}{\sum_i w_{i,RE}^2} + \frac{1}{\sum_i w_{i,RE}})$	reml	–
Rukhin Bayes zero estimator	$\hat{\tau}_{rb0}^2 = \frac{\sum_i (Y_i - \bar{Y}_{FE})^2}{k+1} - \frac{\sum_i (n_i - k)(k-1)\sum_i s_i^2}{k(k+1)\sum_i (n_i - k + 2)}$	rb0	[159]
Rukhin Bayesian positive	$\hat{\tau}_{rbp}^2 = \sum_i (Y_i - \bar{Y}_{FE})^2/(k+1)$	rbp	[159]

A standard confidence interval is calculated as, $\hat{\theta} \pm \hat{\sigma}_\theta \; z_{1-a/2}$, where $z_{1-a/2}$ is the $(1 - a/2)$ quantile of the standard normal distribution and $\hat{\sigma}_\theta = \sqrt{1/\sum_i w_i}$.

To apply the RE model, estimation of heterogeneity is required. In the presence of zero cells, heterogeneity estimators entail the addition of a small continuity correction (CC) on zero cells in order to provide finite estimates. Several methods for estimating τ^2 are proposed in the literature. Table 4.1 presents a summary of the 15 estimators that are included in this study. For a detailed overview of heterogeneity estimators, we refer the reader to two systematic reviews [206, 223].

4.3 Motivating examples

Intravenous immunoglobulin (IVIG) for Guillain-Barre syndrome (GBS)

GBS syndrome has a prevalence of 1-9 /100.000 [224], the term is used to describe a number of rare post-infection neuropathies. Patients may recover completely, remain unable to walk 6 months after disease onset or have a fatal outcome. A recent Cochrane review and MA summarized four RCTs that compared IVIG to plasma exchange [207]. Treatment discontinuation was reported, as a secondary outcome. Trials which were relatively small either failed to report any event or they only had one in each arm. On the contrary, the largest of these trials reported a considerable number of events in both arms (Figure 4.1). For the initial analysis the Mantel-Haenszel (MH) FE risk ratio 0.14 (95% 0.05-0.36) was used. By using the MH, the authors excluded information from trials with no reported event, which may resulted in a significant overall effect with moderate estimated heterogeneity.

Sapropterin dihydrochloride for Phenylketonuria (PHK)

PHK is a common inborn error of amino acid metabolism that causes mental disability (mild to severe) to patients who are not treated properly. It is considered a rare child disorder with a prevalence of 1-5 / 10.000 [224]. A Cochrane review consisted of two studies on sapropterin dihydrochloride and reported on several adverse events, such as vomiting [182]. The two studies produced contradictory but not significant results overall (Supplementary material A - Table 1). Even though, the estimated heterogeneity was substantial, the studies were again pooled using a FE MH on the risk ratio 1.04 (95% 0.28-3.91) [182].

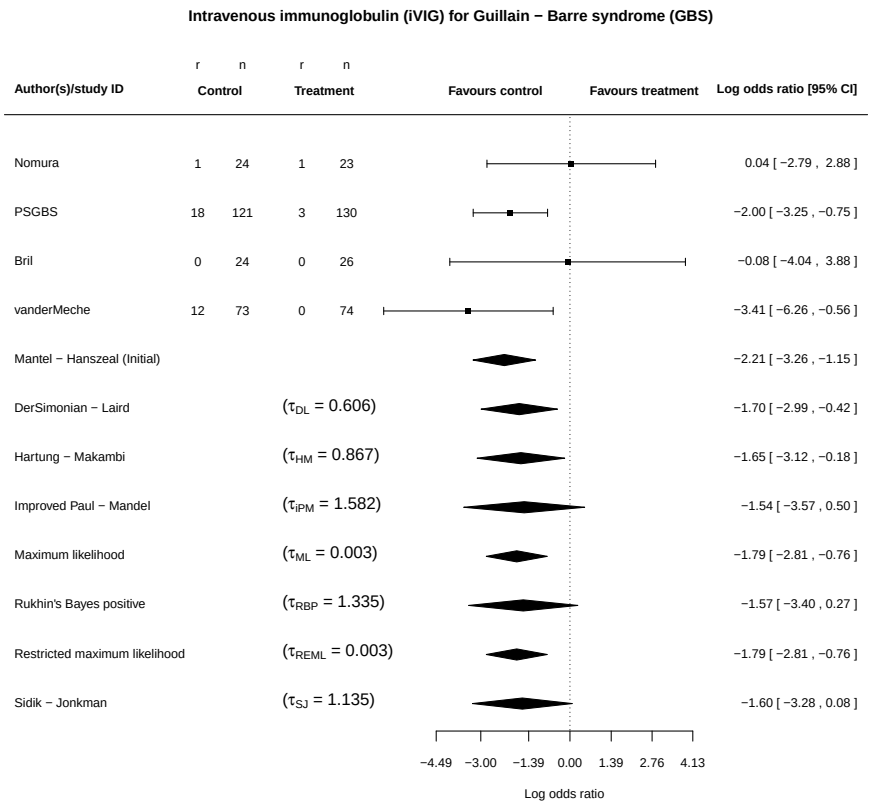

Figure 4.1: Forest plot of the overall treatment effect (log odds ratio) for the Guillain-Barre syndrome (GBS) example. The inverse-variance random-effects method is applied in combination with the seven selected heterogeneity estimators. The confidence intervals are calculated as $\hat{\theta} \pm \hat{\sigma}_\theta \cdot z_{1-\alpha/2}$. The Mantel-Haenszel analysis is plotted as well. For abbreviations see Table 4.1.

Analysis of motivating examples

In regards to our first example (GBS), the final conclusion is influenced considerably by the choice of the heterogeneity estimators. Estimators that lead to a larger estimate value of τ fail to reject the null hypothesis and therefore result in a more conservative conclusion (Figure 4.1).

In the second example (PHK), the overall treatment effect changes direction, depending on the choice of estimator (Supplementary material A - Table 2). The overall treatment effect remains non-significant due to the contradictory results of the two available trials. When estimating the heterogeneity, we observe a behaviour similar to the first example.

4.4 Simulation study

In order to assess the performance of a predefined versus a data-driven choice of analysis in the aforementioned setting, we conducted a simulation study that is divided in two parts; (1) evaluating the operational characteristics for the whole simulation, which represents the "*unconditional approach*" strategy and (2) evaluating the operational characteristics for subsets of the whole simulation that are defined by the number of observed zero cells in a simulated MA. The second part represents the "*conditional approach*" strategy.

Unconditional approach

Following the strategy of Hartung and Knapp [198] we simulated logORs from the null and alternative hypothesis. We varied the overall treatment effect as $\theta \in \{0, 1\}$ and set the heterogeneity equal to $\tau^2 \in \{0, 0.5, 1, 2\}$. These four values correspond to $I^2 \simeq \{0\%, 40\%, 60\%, 75\%\}$ levels of relative heterogeneity, which are calculated via simulation of $I^2 = \tau^2/(\tau^2 + \bar{s}^2)$, $\bar{s}^2 = \sum_{j=1}^{10^5} s_j^2 \, /10^5$ where j: number of simulations. The total number of trials was set within $k \in \{2, 3, 4\}$. Eleven fixed values as of $P_c \in \{0.05, 0.06, ..., 0.15\}$ were used for the control group event rate of the outcome. By simulating a uniformly random draw between $(20, 30)$ for each trial arm, we varied the samples sizes between trials, while we kept the allocation ratio within each trial equal to 1:1. The small sample sizes in combination with different levels of control event rate lead to specific levels of expected zero-event arm percentages (Supplementary material A - Table 3).

Conditional approach

For the second approach we focused on the evaluation of a four (k=4) trial MA, since the relative performance of the heterogeneity estimators was similar across k=2,3,4 trials. The conditional simulation theoretically leads up to a maximum of 9 distinct subsets, since a four

trial MA results from a minimum of 0 to a maximum of 8 zero-event arms. Of course, the latter ones are not useful to consider for a meta-analysis.

For the unconditional approach we based performance measures on 10,000 simulated MAs and evaluated all 15 τ estimators, while for the conditional approach we based performance measures on 1,000,000 simulated MAs and evaluated 7 selected τ estimators that we considered important from the unconditional analysis. A constant CC of 0.5 was added to all cells of a trial that reported at least one zero event. An overview of the varied parameters for each simulation approach is presented in Supplementary material A (Table 4).

Performance measures

We assessed the bias of heterogeneity and overall treatment effect estimates. We calculated the empirical type I error, the power and coverage of the 95% confidence interval of the overall effect estimate. Finally, we computed the probability of each estimator to observe a non-zero heterogeneity estimate ($Pr(\hat{\tau}^2) > 0$) .

4.5 Results

In our small population settings, many heterogeneity estimators performed similarly. More specifically, estimators can be grouped -based on their performance- in two groups. Estimators dl, dl2, dlp, he, he2, mvvc, pm and rb0 displayed similar behaviour in our study. Estimators ml and hs showed a similar insufficient performance in identifying heterogeneity (Supplementary material B). Based on this we selected a key set of 7 estimators for detailed evaluation; dl from the first group, ml from the second group and five estimators that displayed the most divergent behaviour sj, ipm, rbp, hm and reml. In the case of two studies, most heterogeneity estimators behaved similarly.

Regarding the unconditional approach, we summarize the results in two figures Figure 4.2 ($\tau^2 = 0$) and Figure 4.3 ($\tau^2 = 1$). The same two scenarios are presented for the conditional approach in Figure 4.4 and 4.5. Interested readers can find the figures of the remaining scenarios in Supplementary material B.

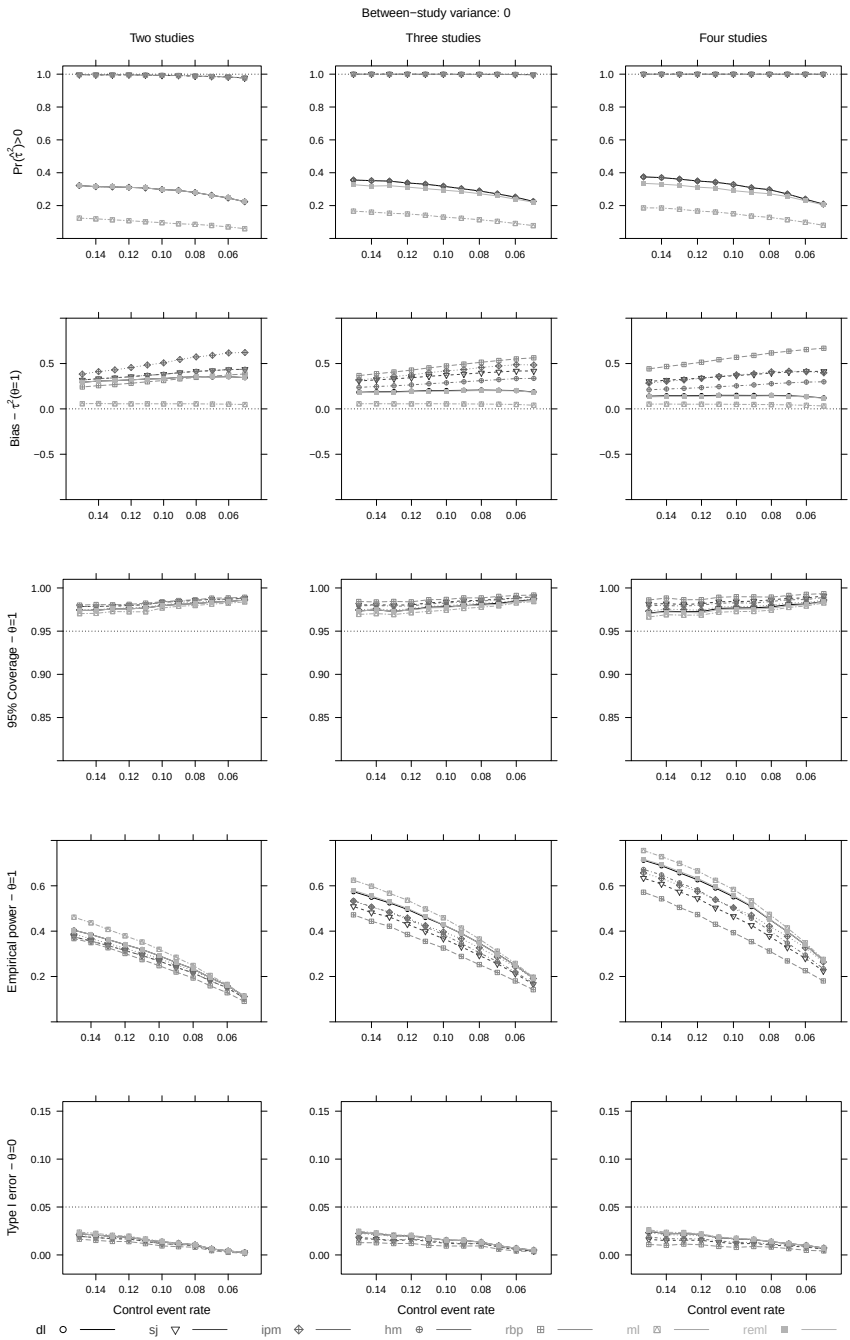

Figure 4.2: Unconditional approach operational characteristics ($Pr(\hat{\tau}^2) > 0$, mean bias of τ, coverage of the 95% confidence intervals, empirical power and type I error of θ) for two to four studies and $\tau^2 = 0$. For abbreviations see Table 4.1.

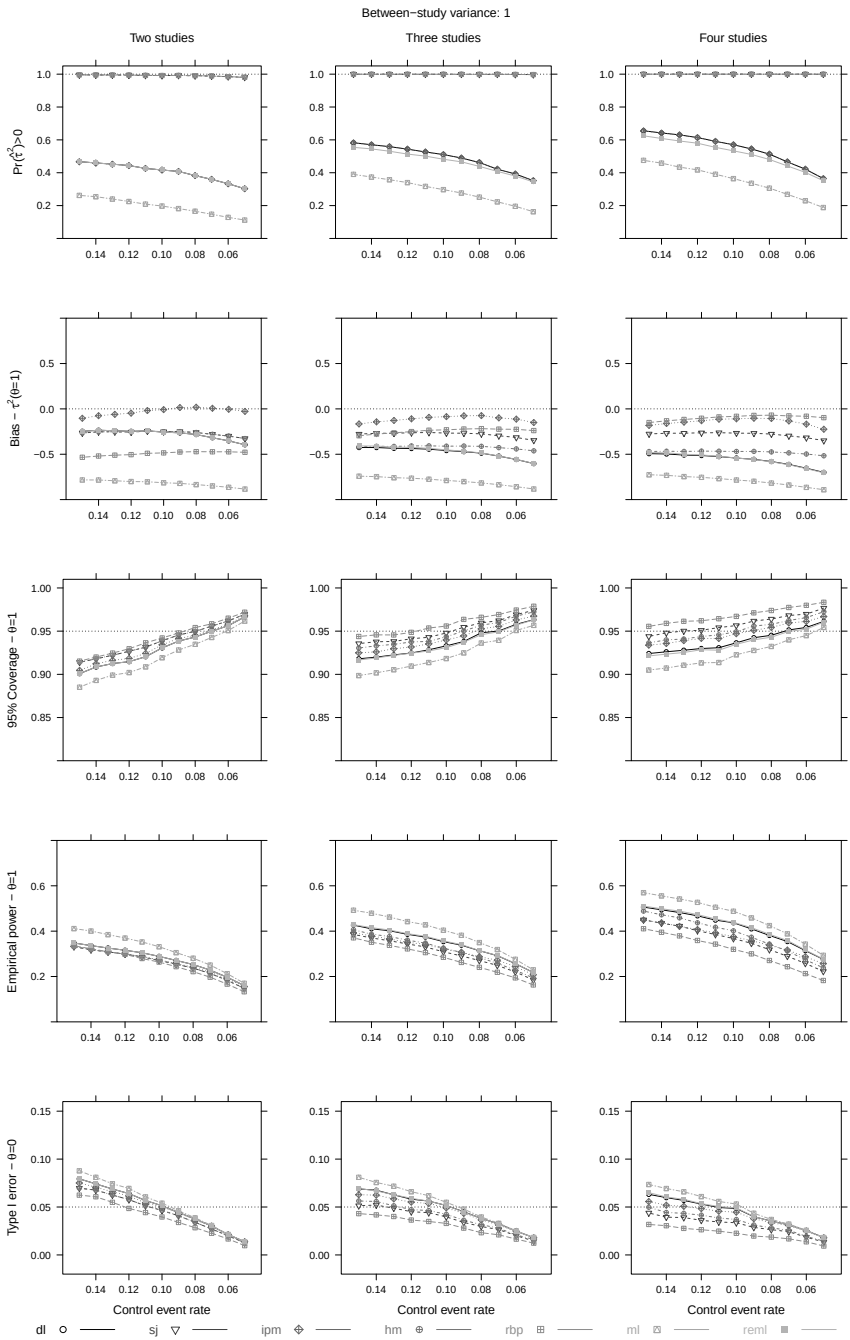

Figure 4.3: Unconditional approach operational characteristics ($Pr(\hat{\tau}^2) > 0$, mean bias of τ, coverage of the 95% confidence intervals, empirical power and type I error of θ) for two to four studies and $\tau^2 = 1$. For abbreviations see Table 4.1.

Unconditional approach

Alternative heterogeneity estimators had little impact on the bias of $\hat{\theta}$. As the control event rate (P_c) decreases, bias increases for all estimators. In addition, the point estimation of τ is problematic as well. Under homogeneity ($\tau^2 = 0$), all estimators greatly overestimate τ, except for ml, while under heterogeneity ($\tau^2 = 1$) rbp, sj and ipm induce the least bias on τ (Figures 4.2 and 4.3).

The presence of heterogeneity impacts the type I error heavily. In non-sparse conditions, when $\tau^2 = 0$, all estimators behave conservatively in interval estimating the overall effect, while in heterogeneous conditions ($\tau^2 = 1$) most of the estimators behave liberally. On the contrary, all estimators display conservative behaviour in very sparse conditions, regardless of the presence of heterogeneity (Figure 4.2 and 4.3). In addition, decreasing P_c levels impact the 95% coverage. We also note that no estimator shows potential to control the coverage, when only two or three small trials are available (Figures 4.2 and 4.3).

The properties of the estimators' depend on the levels of true heterogeneity. As true heterogeneity will not be known, nor very reliably estimated we seek some robustness. And thus, we would prefer estimators that are less dependent on levels of true heterogeneity; for example, sj, hm and ipm (Figures 4.2 and 4.3).

Conditional approach

The first row in Figures 4.4 and 4.5 represents simulations that produce a specific number of zero cells. The first column refers to MAs with no observed zero cell. The rest refer to MAs with an exact number of observed zero cells.

In terms of bias of $\hat{\theta}$, we notice similar properties across conditional subsets; hence, an increase of negative bias, as the P_c decreases (Figures 4.4 and 4.5). In the particular case of exactly no zero cell we observe an overall negative bias (Figures 4.4 and 4.5). The point estimation of τ is impacted by zero cells as well. When no zero cell trial is observed in a MA, all estimators produce values that are relatively close to each other. The increasing number of zero cells makes the estimation of heterogeneity unstable (Figures 4.4 and 4.5).

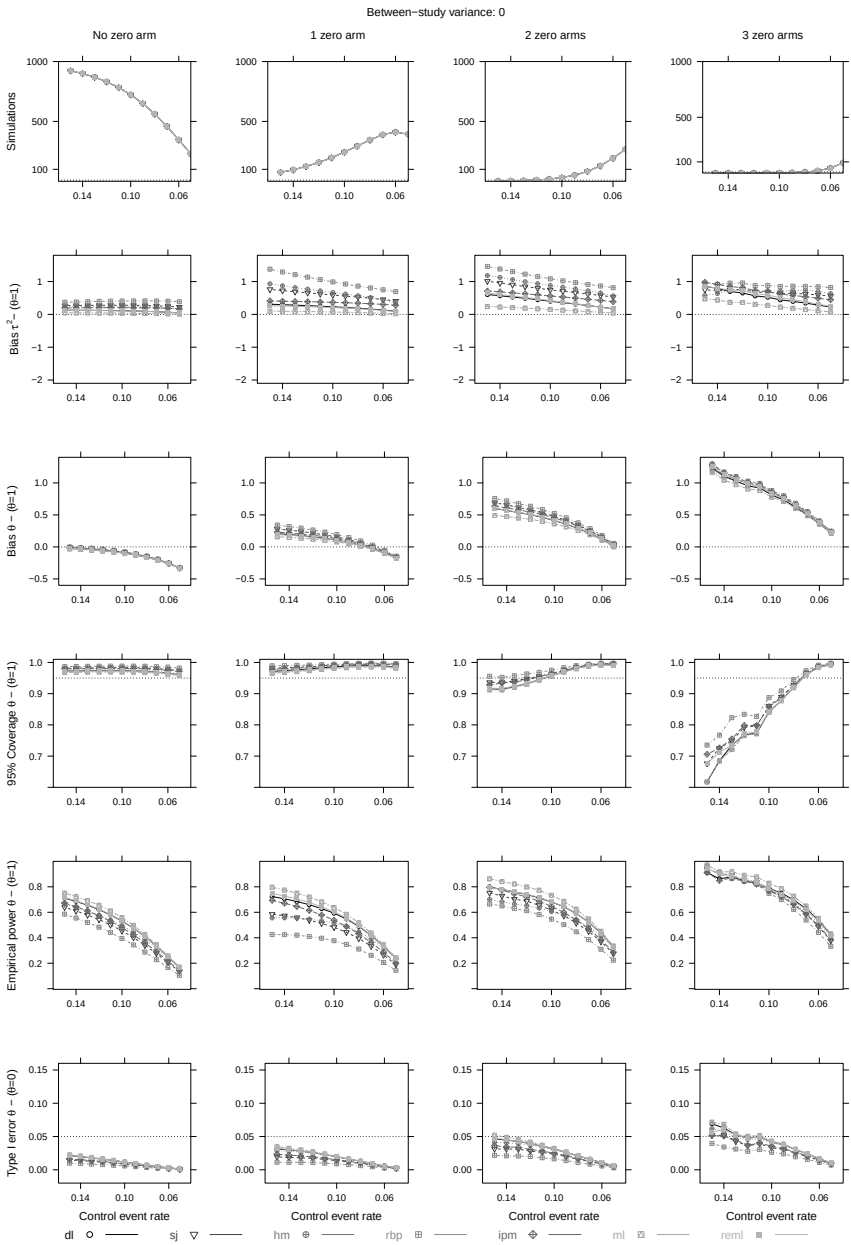

Figure 4.4: Conditional approach operational characteristics ($Pr(\hat{\tau}^2) > 0$, mean bias of τ, mean bias, coverage of the 95% confidence intervals, empirical power and type I error of θ) for four studies and $\tau^2 = 0$. For abbreviations see Table 4.1. First row y-axis - 1000: 1,000,000, 500: 500,000, 100: 100,000 simulations.

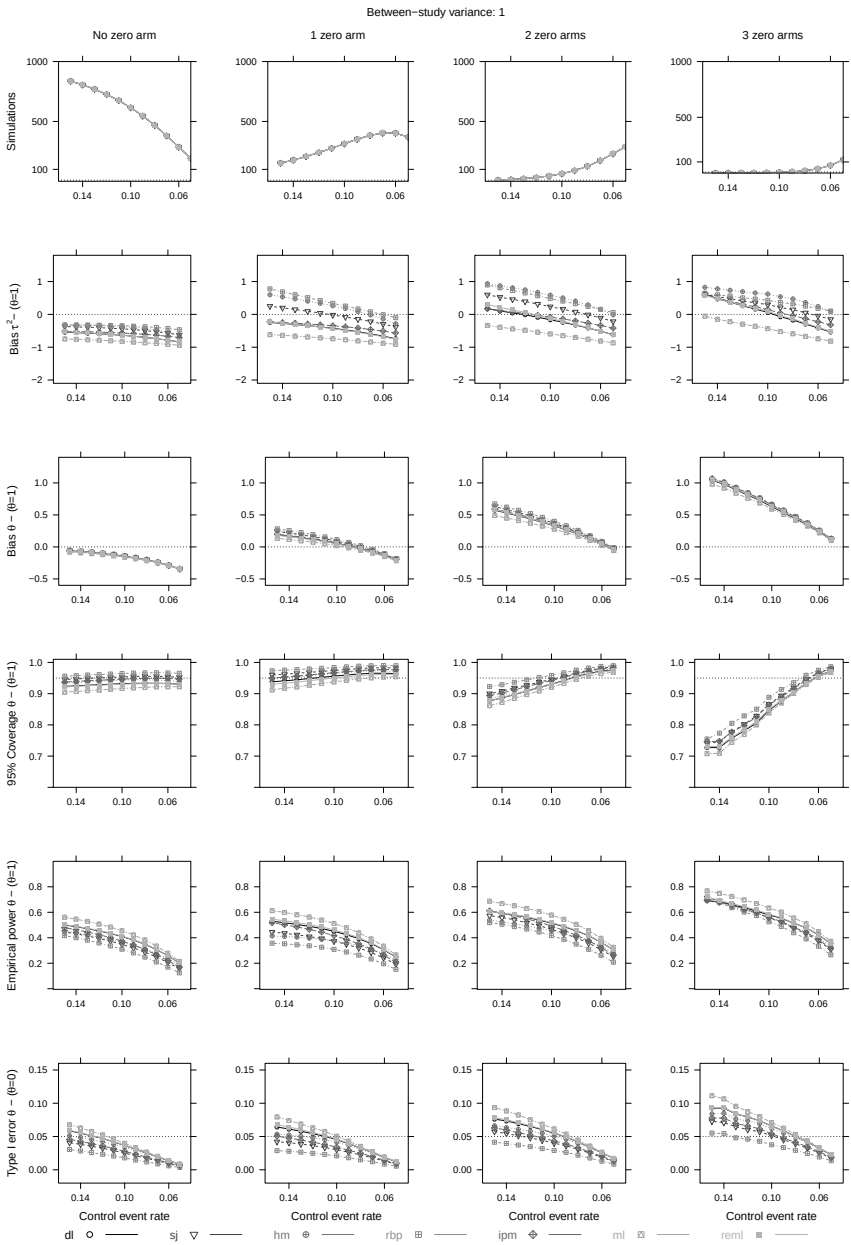

Figure 4.5: Conditional approach operational characteristics ($Pr(\hat{\tau}^2) > 0$, mean bias of τ, mean bias, coverage of the 95% confidence intervals, empirical power and type I error of θ) for four studies and $\tau^2 = 1$. For abbreviations see Table 4.1. First row y-axis - 1000: 1,000,000, 500: 500,000, 100: 100,000 simulations.

The performance of the estimators in terms of 95% coverage and type I error, depends again on the levels of true heterogeneity. In homogeneous cases ($\tau^2 = 0$), independently of observed zero cells, all estimators lead to conservative inferences. When no zero cell trial is observed in a MA, and heterogeneity exists ($\tau^2 = 1$), then most estimators result in liberal inferences. As the number of zero cells increases, estimators result in conservative inferences (Figure 4.5). Again estimators whose performance is less dependent on levels of true heterogeneity are sj, hm and ipm. In addition, ipm produces relative higher power in comparison to sj and hm when one or two zero cells are observed in a MA (Figures 4.4 and 4.5).

In the case of no observed zero cells in a MA of heterogeneous settings ($\tau^2 \geq 1$), all estimators induce negative bias on the estimation of θ and the estimation of τ (Figure 4.5). When at least one zero trial is observed, inference becomes unstable. Such a behaviour could be explained by the impact of CCs on the study weights. When a zero cell trial is observed and a CC is applied, this trial's weight decreases. Therefore, RCTs with low event rates that probably point towards a small or no treatment effect would be down-weighted.

Revisiting the motivating examples

According to our simulation study, the conditional selection of heterogeneity estimator, which is based on the exact number of zero cells, would bring no added value, compared to the unconditional selection of an estimator when a sparse-events MA in small populations is expected. As heterogeneity cannot be reliably estimated in such sparse settings, the chosen estimator should be robust against the level of true heterogeneity. For example, if we had selected the sj, an estimator that was found to be less impacted by the levels of true heterogeneity, we would not have rejected the null hypothesis for the GBS example (Figure 4.1).

Supplementary material A (Table 2) presents an extensive analysis that demonstrates the effect of applying alternative heterogeneity estimators on the overall treatment effect for the two motivating examples.

4.6 Discussion

In this paper we assess and discuss the problematic (interval) estimation of the overall treatment effect, in the presence of heterogeneity for a MA of a few small RCTs with zero events. In this context a truly robust estimation of heterogeneity appears not feasible. Neither can we recommend a single heterogeneity estimator which provides overall satisfactory performance in our small population sparse-event setting. In addition, the comparison between the two simulation approaches showed that the relative performance of heterogeneity estimators did not differ. Therefore, there is no material issue between making a predefined (unconditional) or a data-driven (conditional) choice. Further insights are provided by the conditional approach, which showed that even one observed zero cell has a considerable impact on the inference.

When performing a MA of rare diseases with anticipated or reported zero cells, regardless of a predefined or a data-driven analysis choice, one should avoid methods whose performance depends strongly on the presence of heterogeneity. Following this context, we identify and suggest estimators that perform modestly robust in (interval) estimating the overall treatment effect across a range of heterogeneity assumptions such as sj, hm and ipm. On the contrary, estimators whose performance depends heavily on the true level of heterogeneity, such as rbp and ml, should be avoided. In such a setting, one strategy might be to apply the key set of heterogeneity estimators. If this leads to treatment effect estimates and confidence intervals, which are not comfortably in the same direction, we should probably be cautious to draw firm conclusions.

With few events, the estimated study effects are biased, a bias which reveals itself in between-study variance. Few events also result in large within-study variance which masks between-study variance. Therefore, a trade-off exists; due to the biased effect estimates, heterogeneity increases but due to large within-study variances, heterogeneity decreases. Hence, we conclude the following; (i) when no heterogeneity exists it can only be overestimated due to the biased estimates but (ii) when large heterogeneity exists, it is masked and underestimated.

The simulation study results pair with previous research. In our small population setting, a number of heterogeneity estimators showed small differences in performance [158]. In the

particular case of two studies, most of the heterogeneity estimators behaved similarly as also was theoretically expected [158]. As noted already, a considerable difference was observed on the (interval) estimation of the overall treatment effect among heterogeneity estimators that are known to overestimate (rbp) or underestimate (ml) the true heterogeneity [206, 215, 221]. Such choices should be avoided in our setting as their performance is dependent on the level of true heterogeneity, which cannot be properly estimated.

We only considered a simple Wald test for hypothesis testing via the IV method. We note the existence of an alternative test [198], which has the ability to control the type I error, in a more effective manner than the Wald test for a small number of trials. However, this test does not have sufficient power to detect a true effect [11, 16]. In addition, the simple IV RE model might underperform in a few trials MA, thus sophisticated techniques that control type I error might be preferred. In this context a sensitivity analysis based on a variety of techniques is suggested [225].

Simulation studies have evaluated several other meta-analytical methods regarding their ability to account for zero cells [149, 211, 213, 214]. Among others, they include: (1) the evaluated IV method with alternative CCs [211], (2) the Peto method, which excludes trials with zero events in both arms internally from a MA [214], (3) the MH method for the OR [214], (4) methods that use alternative effect measures, such as the arcsine difference [213] and (5) multilevel models or with alternations in their likelihood [149]. The latter are prone to convergence issues when the number of levels (groups or trials) and the number of events or patients is limited [42, 149]. These studies [149, 211, 213, 214] focused on sparse-events MA, particularly in cases of relatively large sample sizes and large numbers of available studies. Hence, results could not be generalized directly to rare diseases, as the latter have both a limited number of trials and small sample sizes. Further research could focus on the aforementioned methods' behaviour, on the basis of the exact number of observed zero cells in a MA when only a few trials are available.

Further, by utilizing historical data, experts' opinions or priors that cover plausible heterogeneity values, Bayesian inference might provide a suitable alternative for cases of small populations [13, 173, 226]. Although it was not our primary focus, initial evaluations showed that a similar two-level normal Bayesian hierarchical model combined with informative priors

[13] produces smaller biases on the estimation of heterogeneity but similarly problematic 95% coverage for very low control event rates.

In this study, we did not evaluate heterogeneity estimation within complex meta-analytical methods, such as a multiple outcome MA [123] or a network MA [10, 227]. However, we expect that the impact of zero cells in small MAs could be relevant for this context as well, and a similar conditional examination could offer further insight.

Concluding, the choice of heterogeneity estimator does not need to depend on whether or not zero cells are observed in a MA of few small trials. Therefore, regardless of a predefined or data-driven analysis choice, when dealing with zero cells in a MA of rare diseases, we recommend methods with performance that does not strongly depend on the presence or absence of heterogeneity.

Supplementary material and higher resolution images can be found at `http://dx.doi.org/10.1016/j.conctc.2017.11.012 X`

Chapter 5

Prior distributions for variance parameters in a sparse-event meta-analysis of a few small trials

K Pateras
S Nikolakopoulos
KCB Roes

Pharmaceutical Statistics. 2020
https://doi.org/10.1002/pst.2053

Chapter 5

Abstract

In rare diseases, typically only a small number of patients is available for a randomized clinical trial. Nevertheless, it is not uncommon that more than one study is performed to evaluate a (new) treatment. Scarcity of available evidence makes it particularly valuable to pool the data in a meta-analysis. When the primary outcome is binary, the small sample sizes increase the chance of observing zero events. The frequentist random-effects model is known to induce bias and to result in improper interval estimation of the overall treatment effect in a meta-analysis with zero events. Bayesian hierarchical modelling could be a promising alternative. Bayesian models are known for being sensitive to the choice of prior distributions for between-study variance (heterogeneity) in sparse settings. In a rare disease setting, only limited data will be available to base the prior on, therefore, robustness of estimation is desirable. We performed an extensive and diverse simulation study, in terms of prior densities, aiming to provide practitioners with advice on the choice of a sufficiently robust prior distribution shape for the heterogeneity parameter. Our results show that priors that place some concentrated mass on small τ values but do not restrict the density, e.g. the $Uniform(-10, 10)$ heterogeneity prior on the $log(\tau^2)$ scale, show robust 95% coverage combined with less overestimation of the overall treatment effect, across a wide range of heterogeneity levels. We illustrate the results with meta-analyses of a few small trials.

5.1 Introduction

To reach firm conclusions, randomized controlled trials (RCTs) commonly require large enough sample sizes, but this is not always feasible for (very) rare diseases [1] in which the limited patient population leads naturally to small RCTs [100]. In RCTs, dichotomous outcomes are common as they facilitate straightforward clinical interpretation for both efficacy and safety. When combined with small sample sizes and low to moderate event rates, such outcomes lead to a large probability of observing zero events on one or more trial arms.

Even in rare diseases usually more than one trial is available for evaluating a (new) treatment [207, 228]. The small sample sizes make it particularly valuable to pool the data in a meta-analysis (MA). To synthesize available RCTs, the standard random-effects MA model is usually applied, also known as the normal-normal hierarchical model.

When zero events are observed, a complication arises for commonly employed frequentist MA methods. Continuity corrections are needed, usually through adding a constant number to the zero cells. These corrections may affect the study-specific treatment effect estimates and inflate their variances [211]. Kuss evaluated likelihood-based MA methods [149], which incorporate information from trials with zero events in one or both treatment arms without the use of such corrections and showed that these performed adequately in a non-small sample and a sufficiently numbered meta-analysis setting. In a similar setting, either variations on the type of treatment effect measure or the use of the Mantel-Haenszel method has been suggested in previous simulation studies [149, 211, 213, 214].

Bayesian MA methods were shown to perform more robustly in MAs with only a few small trials [13, 147, 148, 229]. When synthesizing conveniently large trials, the choice of prior distributions does not impact inference materially [140, 146, 172, 230, 231]. On the contrary, when pooling a few small trials, only a small number of observations contribute to the model likelihood, therefore, inference becomes prior driven [232]. For the normal-normal hierarchical model, a reference prior was suggested that has the ability to maximize the data impact on inference [148]. Under a normal-normal hierarchical model, the use of priors that cover plausible heterogeneity (τ) ranges has been advocated for a Bayesian MA of a few trials [131, 144, 147]. Such priors may not behave similarly when there are zero events in one or

both arms, and specific choices of prior shapes may be preferable; i.e. according to the way they distribute prior mass across $\tau-$values. The normal-normal hierarchical model has been shown to perform poorly in the presence of zero events in a meta-analysis of rare diseases [150]. The use of different distributional model assumptions such as the binomial-normal hierarchical model may be preferable as (a) it avoids the need for continuity corrections, (b) it directly models the events through a logit link function and (c) it can impose dissimilar baseline effects.

The focus of this paper is to investigate the impact of alternative heterogeneity priors on the (interval) estimation of the overall treatment effect and to provide suggestions for a robust Bayesian MA of a few small sparse-event trials. Robust priors should retain sensible and predictable operational characteristics throughout a range of unknown parameter values. The paper is organized as follows. In section 5.2 we describe a basic Bayesian MA hierarchical model, along with the different priors and prior groups on heterogeneity parameter. Section 5.3 presents two motivating examples and their analysis. In sections 5.4 and 5.5 we describe a simulation study that evaluates the selection of priors. In section 5.6 we revisit the examples. Finally, in section 5.7, we summarize the main findings, while the paper ends with a discussion, as well as recommendations for practitioners.

5.2 Bayesian inference in meta-analysis

Bayesian hierarchical model for meta-analysis

We consider a set of k two-armed RCTs with a binary outcome; patients are randomized over two groups: treatment (T) and control (C) resulting in a 2x2 table (Table 5.1).

Table 5.1: Two way table for notation of the i_{th} trial of a meta-analysis.

	Treatment	Control	Total
Events	r_{iC}	r_{iT}	m_i
Non Events	$n_{iC} - r_{iC}$	$n_{iT} - r_{iT}$	$N_i - m_i$
Total	n_{iC}	n_{iT}	N_i

In each trial $i \in (1, 2, ...k)$ and treatment group $j \in \{C, T\}$, the number of events is modelled to follow a binomial distribution $r_{ij} \sim Binomial(\pi_{ij}, n_{ij})$. By π_{ij} we denote the probability of

an event and by n_{ij} the number of subjects of treatment arm j of trial i [233]. Under a random-effects assumption, a commonly-used Bayesian two-level binomial-normal hierarchical model [132, 234] can be written, using the control group as reference, as follows:

$$\begin{aligned} r_{ij} &\sim Binomial(\pi_{ij}, n_{ij}) \\ logit(\pi_{iT}) &= \mu_i + 0.5 * \delta_i \\ logit(\pi_{iC}) &= \mu_i - 0.5 * \delta_i \end{aligned} \tag{5.1}$$

where $\delta_i \sim N(\delta, \tau^2)$, so that τ^2 denotes the between-study variance and δ_i denotes the study-specific effects of treatment vs. control on the log odds ratio (logOR) scale.

We assume a fixed weakly diffuse normal prior on the overall treatment effect $\delta \sim N(0, 100)$ throughout and a diffuse normal prior on $\mu_i \sim N(\mu_0, 100)$ centred around $\mu_0 = \sum_{i=1}^{k} \mu_i/k$ [173]. In comparison to another common choice of hyper-parameter variance value $\delta \sim N(0, 1000)$, we lowered the assumed prior variance to produce more stable inferences [160]. The chosen prior on δ has a 95% range of (-19.6,19.6) in the logOR scale. The heterogeneity parameter can be modelled through alternative prior distributions so that for a transformation of τ, $g(\tau) \sim f(.)$, where $g(\tau)$ denotes a transformation of τ and $f(.)$ denotes a probability density function.

Priors on heterogeneity

While conducting a meta-analysis, the estimation of heterogeneity is rarely of primary interest. In cases of small and sparse meta-analyses, estimation of τ can quickly become infeasible. Therefore, the choice of heterogeneity priors shall also be driven by its ability to aid the proper estimation of the treatment effect. Different priors have been suggested in the literature, for several functions of τ (Table 5.2). In such sparse settings, the impact and behaviour of each prior is based primarily on its distributional shape. Therefore, a sensible manner of clustering such priors would be to evaluate the way they distribute prior mass on the same scale, i.e. on τ scale. In this context, priors can be clustered in, at least, the following four groups. First, Type A priors place more mass close to 0 but support very large values of τ as well [146, 231] (see Figure 5.1). The $Gamma(\alpha, \beta)$ prior distributions (AG, ag) on the precision ($v_\tau = 1/\tau^2$) and a less restrictive prior on $Uniform(-10, 10)$ on the $log(\tau^2)$ scale (AU) can be gathered in

this category. Type B priors place more mass in larger values of τ; i.e. $Uniform$ on τ^2 scale (C, c). Type C priors place mass uniformly in a selected range of τ (i.e. $Uniform$ on τ scale (B, b)). Finally, Type D priors place most of the mass in small values of τ but naturally bound the range to more plausible values than Type A priors; i.e. $Half\text{-}normal$ priors (DN, dn) on τ and a more informative prior version of $Uniform(-10, 1.386)$ on the $log(\tau^2)$ (du). Type D prior distributions are advocated for MA of a few trials [144, 145, 146, 235]. Within each prior we examine two options based on the informativeness provided by their hyper-parameters, one less restrictive (AG, AU, B, C, DN) and one more restrictive (ag, b, c, dn, du) alternative (Table 5.2).

Table 5.2: Description of considered priors on the heterogeneity τ of a Bayesian meta-analysis. $s_0 = \sqrt{k/\sum(s_i^{-2})}$ and s_i^2 are the within-study variances. ID - Abbr. : Identification letter and abbreviation for each prior.

ID - Abbr.	$g(\tau)$	$\sim f(.)$	Restrictive	τ Median	τ (95% range)
AG	$1/\tau^2$	$\sim Gamma(0.001, 0.001)$	Less	> 100	$(> 100, +\infty)$
ag	$1/\tau^2$	$\sim Gamma(0.1, 0.1)$	More	0.3	$(12.9, > 100)$
AU	$log(\tau^2)$	$\sim Uniform(-10, 10)$	Less	1	$(0.01, > 100)$
du	$log(\tau^2)$	$\sim Uniform(-10, 1.386)$	More	0.1	$(0.01, 1.7)$
B	τ^2	$\sim Uniform(0, 1000)$	Less	22.4	$(5, 31.2)$
b	τ^2	$\sim Uniform(0, 4)$	More	1.4	$(0.3, 2)$
C	τ	$\sim Uniform(0, 100)$	Less	50	$(2.5, 97.5)$
c	τ	$\sim Uniform(0, 2)$	More	1	$(0.05, 1.95)$
DN	τ	$\sim Half\text{-}normal(0, 100)$,	Less	6.75	$(0.3, 22.4)$
dn	τ	$\sim Half\text{-}normal(0, 1)$,	More	0.7	$(0.03, 2.24)$
E	$s_0/(s_0 + \tau)$	$\sim Uniform(0, 1)$	–	–	–
e	τ^2	$\sim Half\text{-}normal(0, \Phi(0.75)/s_0)$,	–	–	–

Finally, we use the estimates of the within-study variances (s_i^2) to examine two data-driven priors (E, e) that both incorporate the harmonic mean ($s_0 = \sqrt{k/\sum(1/s_i^2)}\ , i = 1, 2, ...k$) of the s_i^2 of the trials included in the MA [131, 236]. More specifically, prior E, also known as the $DuMouchel$ prior has been suggested for very small sample sizes and, by utilizing s_0, it induces shrinkage on the τ prior distribution [237]. Small values of s_0 result in a narrow-tailed prior distribution on τ and more shrinkage, while large values of s_0 result in a wide-tailed prior distribution on τ and less shrinkage.

In the following section we introduce two motivating examples, illustrate the results when different priors are used and discuss the implications.

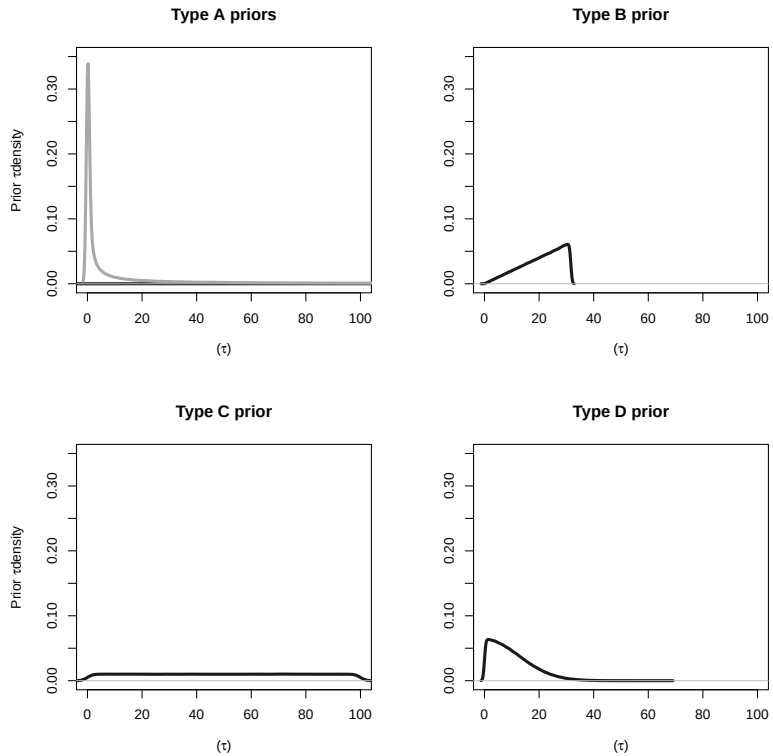

Figure 5.1: Prior density shape considered. The less restrictive options per prior are presented in this figure. Type A priors include both $Gamma$s on v_τ (AG, ag) and the less restrictive $Uniform(-10, 10)$ on $log(\tau^2)$ (AU). The $Gamma$ prior has a very small peak near zero, while the peak of the $Uniform$ type A prior is higher both support very large τ-values. Type B priors include, while both $Uniform$ on τ^2 (B, b), Type C priors include the $Uniform$s on τ priors (C, c), Type D include both the $Half\text{-}normal$ on τ priors (DN, dn) and the more informative $Uniform(-10, 1.386)$ on $log(\tau^2)$ prior (du). All other more informative options within each prior, except for the $Uniform(-10, 1.386)$ on $log(\tau^2)$, retain a similar shape but cover a smaller range of values. The latter more informative prior retains a form closer to Type D priors. For clarity of results the $x - axis$ is graphically truncated for values larger than 100. Figure 3 in Supplementary material I provides a comparison between the less and more restrictive prior.

5.3 Motivating examples

Multifocal motor neuropathy is a progressive rare disorder in which the muscles weaken gradually. Multifocal motor neuropathy is not often fatal but can lead to a significant degree of disability for the patient. Prevalence is estimated at 1-2 cases per 100,000 [238]. A literature review and MA assessed the efficacy and safety of intravenous immunoglobulin in multifocal motor neuropathy [228]. The same evidence was presented in the European Medicines Agency Public Assessment Report of Kiovig [239]. The primary outcome was the improvement in disability scale using MRC (Medical Research Council) scores that evaluate the muscle strength. Three two-arm studies reported the outcome, accounting for a total of 36 recruited patients with 7 reported events in the intravenous immunoglobulin arm and 2 in the placebo arm. The original MA reported no heterogeneity [228].

For the second example, we consider Guillain-Barre syndrome with a MA of four available studies. Guillain-Barre syndrome has a prevalence of 1-9 cases per 100,000 [238] and refers to a number of rare post-infection neuropathies. A literature review and MA summarized RCTs that compared intravenous immunoglobulin to control (plasma exchange) [207]. For one of the secondary outcomes, treatment discontinuation, a few arms reported zero events. This example has been used for evaluating a number of heterogeneity estimators under the inverse-variance method and has been shown to produce conflicting inferences [150] (Chapter 4). Data for both examples are illustrated in Table 5.3.

Analysis of motivating examples

A robust choice of prior is not trivial for our examples. To examine the behaviour of the priors, we use Rjags [185, 240] to fit 3 chains of 850,000 samples after a burn-in of 150,000 samples and a thinning interval of 35 samples for each model. Figure 5.2 presents the posterior median (as a point estimate) and credible intervals of δ and τ for the two motivating examples under different priors. The letters in Figure 5.2 correspond to the letters in Table 5.2.

Table 5.3: Motivating examples; (a) Efficacy endpoint: Improvement in disability, Therapy: Intravenous immunoglobulin vs Placebo, Condition: Multifocal motor neuropathy **(b)** Efficacy endpoint: Treatment discontinuation, Therapy: Intravenous immunoglobulin vs Plasma Exchange, Condition: Guillain-Barre syndrome. $r_{i,j}$ event in control / treatment group, $n_{i,j} - r_{i,j}$ non-event in control / treatment group $\hat{\pi}_{i.}$ = Observed probability of event in each trial, $w_{i,in}$ = Weight of initial analysis.

(a) Multifocal motor neuropathy - Improvement in disability [228]

Author	r_{iT}	$n_{iT} - r_{iT}$	r_{iC}	$n_{iC} - r_{iC}$	$\hat{\pi}_{i.}$	$w_{i,in}$
Azulay	0	5	0	5	0	–
Berg	3	3	0	6	0.25	0.20
Lger	4	3	2	5	0.43	0.80

(b) Guillain-Barre syndrome - Treatment discontinuation [207]

Author	r_{iT}	$n_{iT} - r_{iT}$	r_{iC}	$n_{iC} - r_{iC}$	$\hat{\pi}_{i.}$	$w_{i,in}$
Meche	0	74	12	61	0.08	0.39
Bril	0	26	0	24	0	–
PSGBS	3	127	18	103	0.09	0.58
Nomura	1	22	1	23	0.04	0.03

The choice of prior for τ has substantial impact on the posterior credible intervals for δ. The posterior median for δ varies substantially as well. More specifically, in the multifocal motor neuropathy example, the posterior median δ has a range of (2.31, 3.27) depending on the τ prior choice (Figure 5.2a). In the Guillain-Barre syndrome example, the posterior median δ has a range of (-2.52, -2.80) (Figure 5.2b). The posterior mean of δ in both examples shows even greater diversity. Interval estimation of δ also varies substantially. Different priors and types of priors lead to considerably divergent inference (Figure 5.2). All Type A priors show a similar behaviour upon the estimation of δ in both examples.

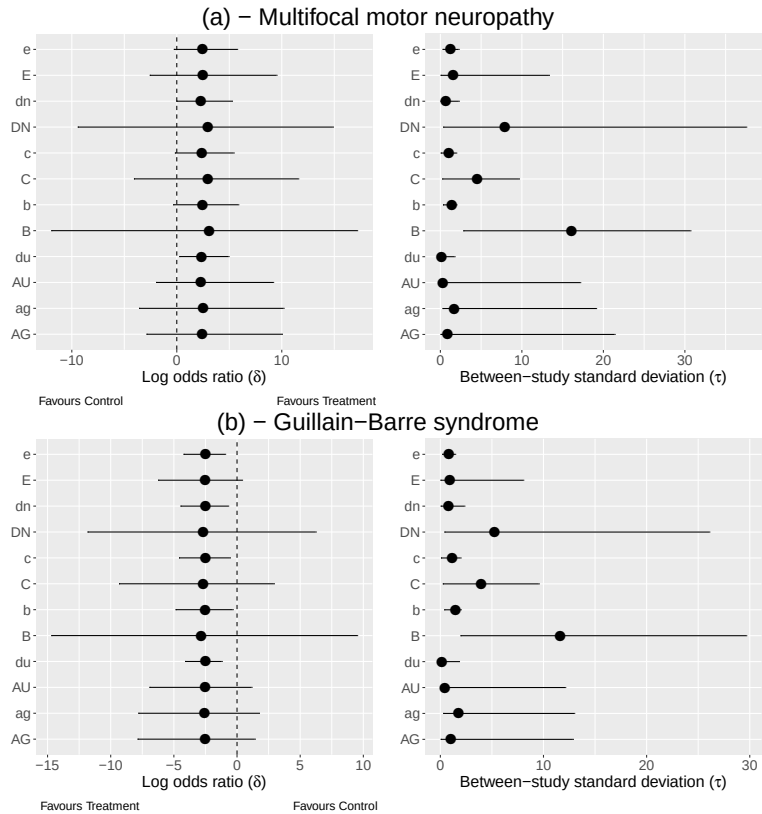

Figure 5.2: Posterior medians and 95% credible intervals of the overall effect (log odds ratio) and the between-study standard deviation (τ) for the two motivating examples (a) Multifocal motor neuropathy and (b) Guillain-Barre syndrome. (AG, ag) - $Gamma$ on v_τ, (AU, du) - $Uniform$ on $log(\tau^2)$, (B, b) - $Uniform$ on τ^2, (C, c) - $Uniform$ on τ, (DN, dn) - $Half\text{-}normal$ on τ, (e) $Half\text{-}normal$ on τ^2, (E) - $DuMouchel$ prior. (AG, AU, B, C, DN) are less restrictive priors on τ and (ag, dn, b, c, dn) are more informative priors on τ.

5.4 Simulation study

To incorporate heterogeneity successfully in both study arms, we simulated study-specific logits for each arm, following the simulation strategy of Hartung and Knapp ([198] , *pRandom* in [241]). Hence, we assumed an initial fixed event probability in the control group and we calculated the event probability in the treatment group, based on a true overall treatment effect. Further, we simulated study-specific logits from a normal distribution with between-

study standard deviation equal to $\tau/\sqrt{2}$ for the control and treatment arm. We utilized the simulated logits to compute the study-arm event probabilities by back-calculating and finally we simulated events for each study arm [241].

We evaluated a number of scenarios by varying the number of trials (k), the number of patients per trial arm (n_{ij}), the control event rate (π_c), the between-study standard deviation (τ) and the overall treatment effect (δ). More specifically, the number of trials varied as $k \in \{2, 4, 6\}$ while patients per trial arm were assumed equally allocated ($n_{iC} = n_{iT}$) and uniformly sampled either between 40 to 50 or between 5 to 10. These sample sizes were selected to represent realistic scenarios for efficacy and safety endpoints of rare and ultra-rare diseases [100]. The control event rate (π_c) in each trial took set values as follows; very low event rate (0.05), low event rate (0.1), moderate event rate (0.3). Specific combinations of sample size and control group event rates lead to particular percentages of zero-event trials in MAs of the simulated data (Supplementary material I - Table 1). The between-study standard deviation took values between $\tau \in \{0.01, 0.5, 1\}$. Finally, we examined three values for the overall treatment effect on the logOR scale, $\delta \in \{0, 0.5, 3\}$.

First, the 12 clustered priors above are evaluated for all scenarios and then a number is selected for further evaluation. Therefore, the number of scenarios is in total 1,994. For each scenario we generated 1,000 simulated datasets. We performed simulations using JAGS [185] and R [242] via a High Performance Cluster. We fitted every model via three parallel chains of 30,000 samples, a burn-in of 4,500 samples and a thinning interval of 5 samples.

In sparse settings the parameters' Markov chain Monte Carlo sampling convergence is of concern. We conducted selective convergence checks on the Markov chain Monte Carlo algorithms via trace plots, convergence diagnostics via the CODA package [243] and focused on the most extreme scenarios of sparsity. We fitted every model via 3 parallel chains and we accounted for autocorrelation by applying a thinning interval of 5 samples. Overall, convergence was achieved. We analytically report on diagnostics in the Supplementary material III, where we compare the convergence of different priors. Diagnostic assessment was performed for both the examples (via generation of 1,000,000 Markov chain Monte Carlo samples) and the simulation study (via generation of 34,500 Markov chain Monte Carlo samples).

Each scenario was mainly evaluated by the following performance measures: (1) average posterior median for δ, (2) coverage of the 95% credible interval (CrI). We also discuss the mean square error of δ and the average posterior median estimates of τ for exploratory purposes and completeness. Prior robustness was defined by the adequate overall measures and small observed fluctuations in coverage of the 95% credible interval among the scenarios considered.

5.5 Results of simulation study

For relatively large sample sizes and higher π_c, regarding the posterior estimation of δ, all priors perform similarly (Figures 5.3 and 5.4). The overall priors performance deteriorates at a low control group event rate ($\pi_c = 0.05$) for a few small RCTs MA, as the average posterior median of δ is overestimated (Figures 5.3 and 5.4) at all levels of true heterogeneity. Furthermore, we observe an overall positive bias in the posterior median estimation of δ, when δ is large.

All Type A priors retain more robust 95% coverage in comparison to other prior groups (Figures 5.3 and 5.4). More specifically, the $Uniform(-10, 10)$ on $log(\tau^2)$ scale prior (AU) retains a more robust 95% coverage at small values of π_c, independently of sample size and it properly estimates the posterior median logOR on average as well (Figures 5.3 and 5.4). The $DuMouchel$ empirical prior (E) shows a comparable behaviour. The 95% coverage of Type B, C and D priors varies throughout the evaluated scenarios from conservative in larger sample sizes to liberal 95% coverage in smaller sample sizes (Figures 5.3 and 5.4). All priors encounter issues regarding the 95% coverage when the treatment effect is large ($\delta = 3$), the sample size is limited and the control event rate very small ($\pi_C = 0.05$) (Figure 5.4).

More informative priors for τ (b, c, dn, du) tend to produce a less variant posterior point estimate of δ, while less restrictive priors that mostly support larger values for τ (B, C, DN) tend to overestimate δ heavily (Figures 5.3 and 5.4). Moreover, the use of the latter group of priors at any level of π_c results in conservative inference for δ (Figures 5.3 & 5.4). This set of less restrictive priors and (du), a prior that also has the smallest prior τ median of all selected (Table 5.2), these four priors performed poorly in terms of 95% coverage irrespective of the sparseness of events.

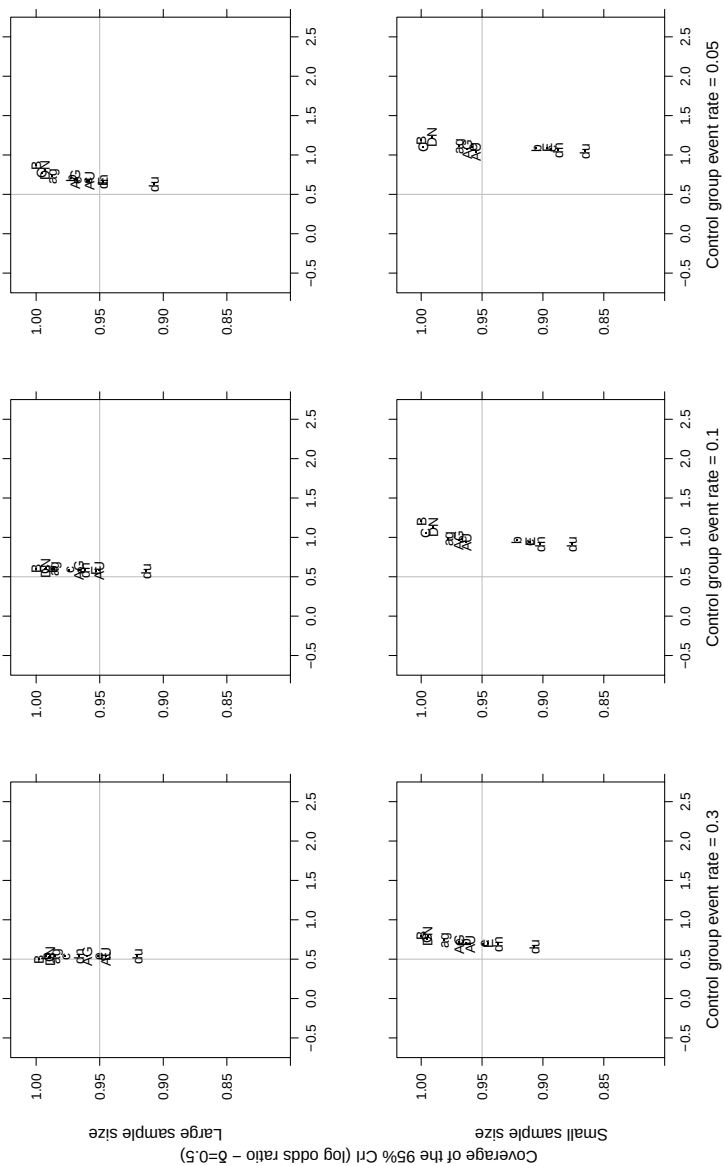

Figure 5.3: Scatter plot of average posterior median overall effect (log odds ratio) against its mean coverage of the 95% CrI for all simulated scenarios (Overall effect: $\delta = 0.5$, between-study standard deviation: $\tau \in \{0.01, 0.5, 1\}$, number of trials: $k \in \{2, 4, 6\}$) of a meta-analysis with control group event rate: $\pi_c \in \{0.05, 0.1, 0.3\}$ with small sample size trials ($n_{ij} \sim Uniform(5, 10)$) or large sample sized trials ($n_{ij} \sim Uniform(40, 50)$). (AG, ag) - $Gamma$ on v_τ, (AU, du) - $Uniform$ on $log(\tau^2)$, (B, b) - $Uniform$ on τ^2, (C, c) - $Uniform$ on τ, (DN, dn) - $Half\text{-}normal$ on τ, (e) $Half\text{-}normal$ on τ^2, (E) - $DuMouchel$ prior. (AG, AU, B, C, DN) are less restrictive priors on τ and (ag, dn, b, c, dn) are more informative priors on τ.

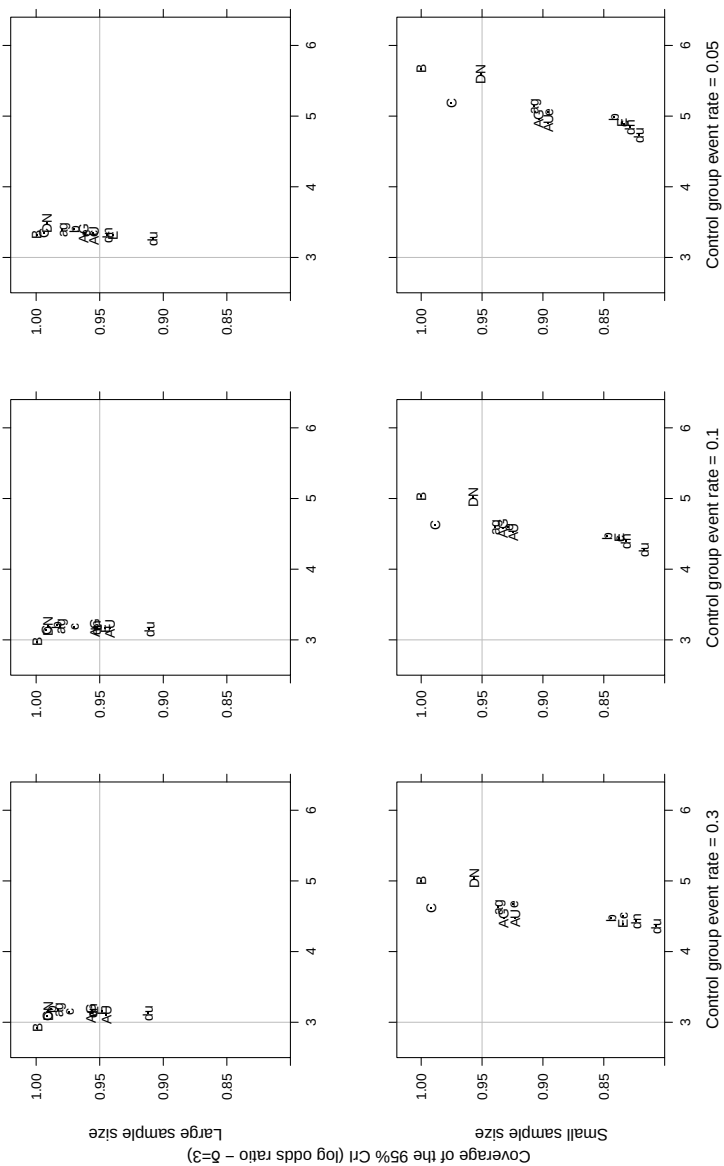

Figure 5.4: Scatter plot of average posterior median overall effect (log odds ratio) against its mean coverage of the 95% CrI for all simulated scenarios (Overall effect: $\delta = 3$, between-study standard deviation: $\tau \in \{0.01, 0.5, 1\}$, number of trials: $k \in \{2, 4, 6\}$) of a meta-analysis with control group event rate: $\pi_c \in \{0.05, 0.1, 0.3\}$ with small sample size trials ($n_{ij} \sim Uniform(5, 10)$) or large sample sized trials ($n_{ij} \sim Uniform(40, 50)$). (AU, Au) - $Gamma$ on v_τ, (AU, du) - $Uniform$ on $log(\tau^2)$, (B, b) - $Uniform$ on τ^2, (C, c) - $Uniform$ on τ, (DN , dn) - $Half\text{-}normal$ on τ, (e) $Half\text{-}normal$ on τ^2, (E) - $DuMouchel$ prior. (AG, AU, B, C, DN) are less restrictive priors on τ and (ag, dn, b, c, dn) are more informative priors on τ.

It should be noted that even though Figures 5.3 and 5.4 only provide a general view of all simulated scenarios, we did not observe deviations regarding the average posterior median of the overall treatment effect (δ) when investigating specific scenarios. Additional averaged and scenario-specific simulations are presented in Supplementary material II.

For clarity of results, after studying all priors (Figures 5.3, 5.4 and Supplementary material II), we focus on four priors (AG, AU, dn, E) which either (1) performed more robustly in the current simulation study (AU, E), (2) are commonly used in the literature (AG) and/or (3) have been suggested in recent literature for meta-analysis of rare diseases (dn) [13]. We present selected scenarios for $\delta = 3$ in the main manuscript (Figures 5.5 and 5.6).

Coverage of the 95% CrI for the overall treatment effect (δ)

The value of the treatment effect does not heavily affect the coverage of the 95% CrI. Specifically, for a MA of four trials, most robust coverage is generally produced by the two Type A priors, the $Gamma(0.001, 0.001)$ prior on v_τ (AG) and (AU), alongside with (E) empirical prior (Figure 5.5). However, in a MA of less than four trials, prior (AG) induces systematically larger deviations from the nominal 95% coverage in comparison to priors (AU) and (E) (Figure 5.5). The Type D $Half\text{-}normal(0, 1)$ prior (dn) prior either induce (1) overcoverage for low levels of true heterogeneity ($\tau \leq 1$) or low event rates or (2) large undercoverage for large true heterogeneity ($\tau = 1$), regardless of the event rate. In comparison to the three priors described above the (dn) prior, show the least robust coverage throughout all scenarios and more particularly for varying sample sizes or levels of τ (Figure 5.5 and Supplementary material II).

Mean square error of the overall treatment effect (δ)

All priors produce comparable levels of mean square error (Supplementary II - Figures 1-3 and 12-15). The priors that produce the least optimal and most divergent behaviour in comparison to the rest are the (DN) and (B) priors.

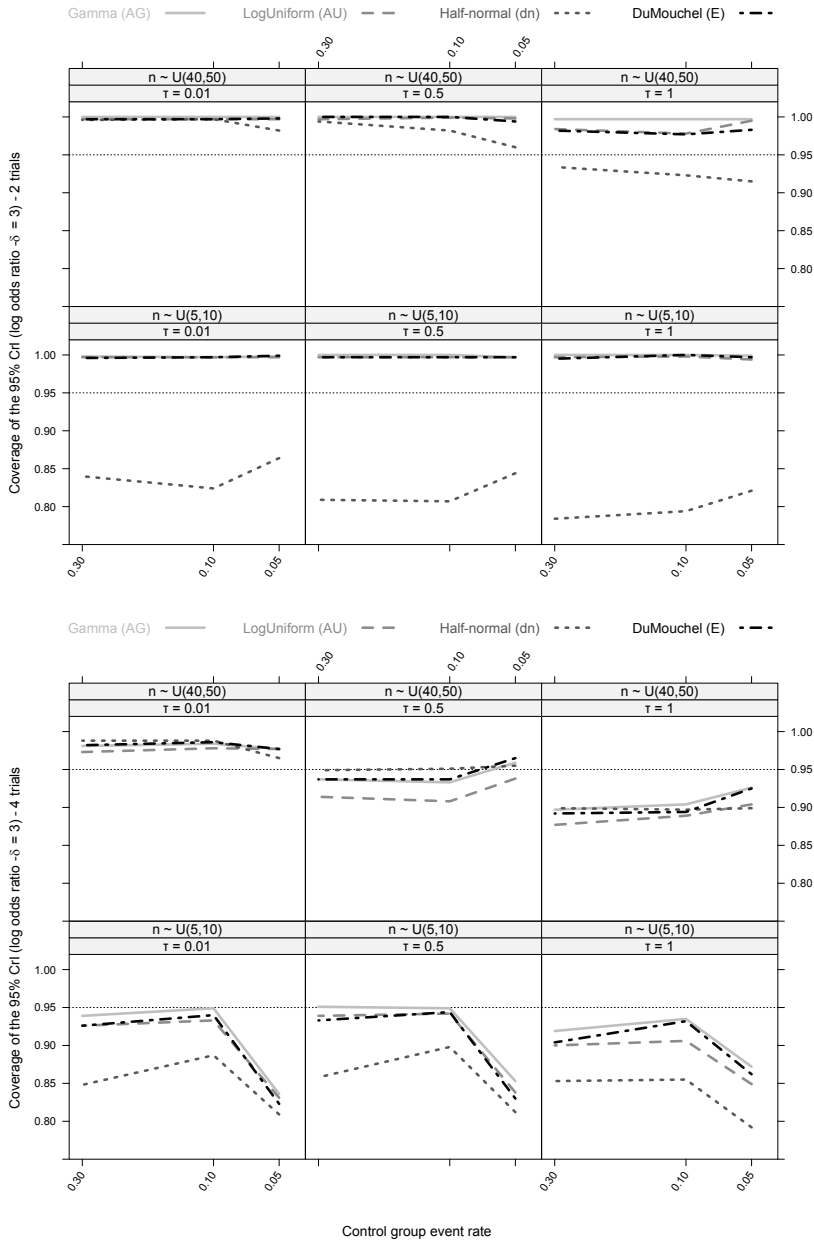

Figure 5.5: Coverage of the 95% CrI line plots of the overall effect (log odds ratio) on different control group event rate levels for a large true overall effect ($\delta = 3$), three values of $\tau \in \{0.01, 0.5, 1\}$ and small sample size trials ($n_{ij} \sim Uniform(5, 10)$) or large sample sized trials ($n_{ij} \sim Uniform(40, 50)$). (AG): $Gamma(0.001, 0.001)$ on v_τ, (AU): $Uniform(-10, 10)$ on $log(\tau^2)$, (dn): $Half\text{-}normal(0, 1)$ on τ, (E): $DuMouchel$ prior. Results for 6 trials can be found in Supplementary Material II.

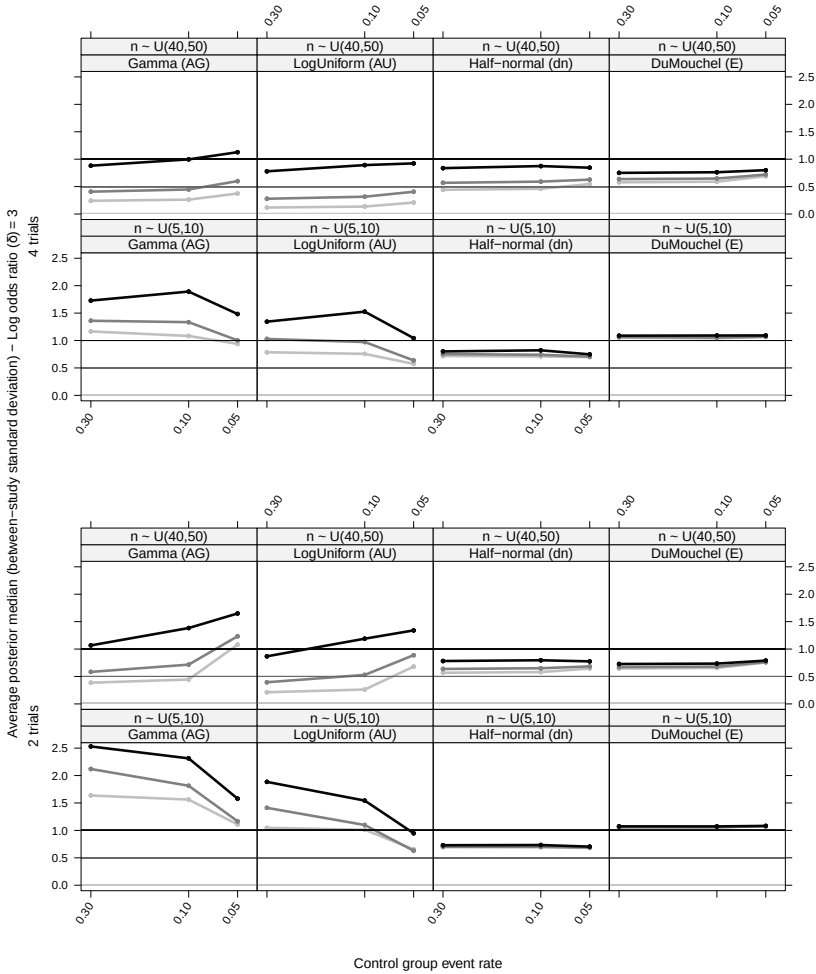

Figure 5.6: Average posterior median line plots of the between-study standard deviation (τ) on different control group event rate levels for a large true overall effect ($\delta = 3$) and small sample size trials ($n_{ij} \sim Uniform(5, 10)$) or large sample sized trials ($n_{ij} \sim Uniform(40, 50)$). The grey lines represent 3 levels of heterogeneity, namely, light grey: $\tau = 0.01$, grey: $\tau = 0.5$, dark grey: $\tau = 1$. (AG): $Gamma(0.001, 0.001)$ on v_τ, (AU): $Uniform(-10, 10)$ on $log(\tau^2)$, (dn): $Half\text{-}normal(0, 1)$ on τ, (E): $DuMouchel$ prior. Results for 6 trials can be found in Supplementary Material II.

Exploring the heterogeneity estimate behaviour (τ)

All 12 priors produced biased results. In less sparse scenarios ($n_{ij} \sim U(40, 50)), k = 4, 6$) the type A priors (AU) and (AG) show the least bias on τ, irrespective of the true heterogeneity level (Figure 5.6 & Supplementary material II). Prior (dn), behaved similarly to all other more informative prior choices and showed difficulty in identifying any level of true heterogeneity (Figure 5.6).

5.6 Revisiting the motivating examples

Following the results of the simulation study, prior type A $Uniform(-10, 10)$ on the $log(\tau^2)$ prior (AU) is preferred for the Guillain-Barre syndrome example (4 trials, low event rates, relatively large sample size). When we apply this prior, the primary inference of these studies would produce a posterior probability of $\delta > 0$ equal to 96%. This is less than the 99% posterior probability which is produced by the Type D $Half\text{-}normal(0, 1)$ prior (dn) on τ, a prior that showed non robust overall but sufficient coverage at low to moderate π_c combined with low to moderate τ settings (Figure 5.5). Therefore, inference with both priors suggests efficacy of intravenous immunoglobulin in comparison with plasma exchange in terms of treatment discontinuation and result in comparable posterior distributions (Figure 5.7) and medians for the logOR ($\delta_{AU} = -2.51$ - $\delta_{dn} = -2.49$).

Likewise, for the more sparse multifocal motor neuropathy example (3 trials, moderate event rates, relatively small sample size), prior (AU) would also be preferred. When we apply this prior, the primary inference for these studies would produce a posterior probability of $\delta > 0$ equal to 93%, but when prior (dn) is applied, the posterior probability becomes 97%, which would have overstated our confidence in the effectiveness of intravenous immunoglobulin regarding improvement in MRC scale, based on results of the simulation study. Similarly to the Guillain-Barre syndrome case study, relying on priors (AU) or (dn) produces comparable posterior median logORs ($\delta_{AU} = 2.32$ - $\delta_{dn} = 2.31$), as expected by the reported simulation study (Figures 5.3 and 5.4).

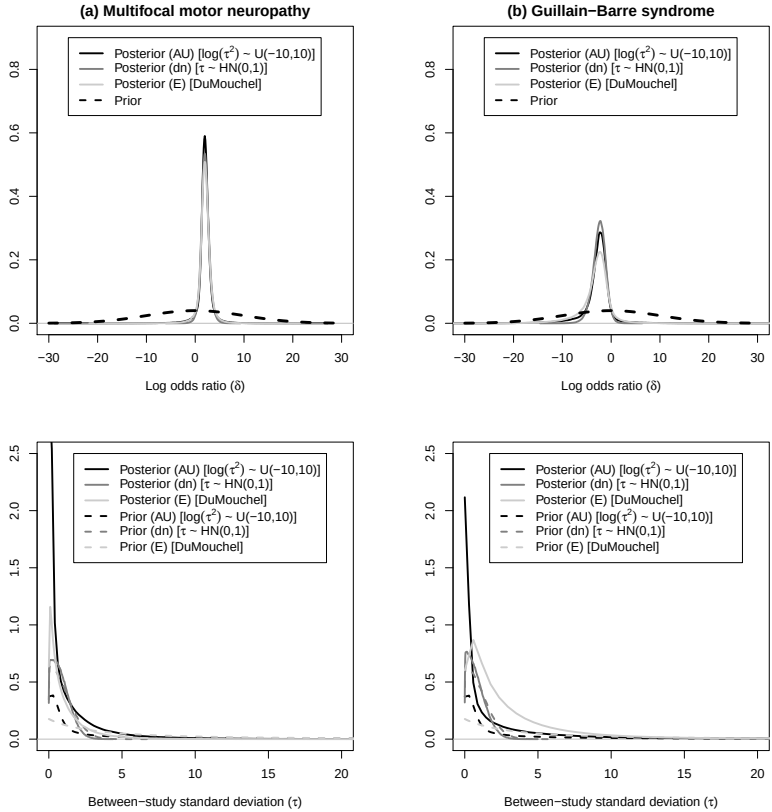

Figure 5.7: Posterior summaries for the overall effect (δ) and the between-study standard deviation (τ) of the Multifocal motor neuropathy and Guillain-Barre syndrome examples for (AU): $Uniform(-10, 10)$ on $log(\tau^2)$, (dn): $Half\text{-}normal(0, 1)$ on τ and (E): $DuMouchel$ empirical prior based on 850,000 iterations with a burn-in of 150,000 iterations and a thinning interval of 35 iterations.

In both examples, data-driven prior (E) produces similar probability statements and posterior median logORs to the Type A (AU) prior, a behaviour which is aligned with the results of the simulation (Figure 5.5). Based on the simulation study, a Type A prior (i.e. AU) that showed robust 95% coverage should be chosen as it provides less variable behaviour in comparison to the studied alternatives under both known and unknown parameters (Types B, C and D), as well.

A comparison between the two priors that performed robustly through the simulation study

(AU and E) and the commonly used half-normal prior (dn) is presented in Figure 5.7 for the multifocal motor neuropathy and Guilen-Barre syndrome examples respectively. In both examples, when prior (AU) is applied, the posterior distribution of τ differs considerably from its prior. However, when prior (dn) is applied, the posterior distribution of τ becomes more prior-driven, a behaviour which is also depicted in our simulation (Figure 5.6). In Supplementary Material I (Table 2), interested readers can find the extended results of all considered prior choices.

5.7 Main findings

(i) The choice of type of prior and prior distribution for τ heavily influences not only the posterior mean/median estimates of τ but also the posterior mean/median estimates of δ in a sparse-events MA of a few small trials.

(ii) In a sparse meta-analysis of a few small ($n_{ij} \sim (5, 10)$) studies, priors that place most of the mass in small values of τ but naturally restrict the range to more plausible values (D) (i.e. dn, du) should be avoided as they do not provide robust point and proper interval estimation of δ.

(iii) Type A priors that place more mass on small values without excluding very large τ prior values (AU) are suggested as a robust choice for a sparse-events MA of a few small trials.

(iv) In many scenarios and even for very sparse settings, the Type A prior $Uniform(-10, 10)$ on the $log(\tau^2)$ scale prior (AU) shows good coverage overall combined with less overestimation of δ or τ in comparison to other prior choices. The $DuMouchel$ prior (E) shows a similar behaviour.

(v) The less restrictive prior choices of priors that place mass uniformly in a selected range (B) and/or priors that place more mass in larger values of τ (C) and the empirical prior (DN) are not appropriate for a sparse-events MA of a few small trials, as they overestimate τ and produce conservative inferences, while resulting in improper estimation of δ. Their more informative alternatives (b, c and dn) produce more reliable inferences at high π_c, but they result in liberal inferences when combined with large true heterogeneity ($\tau = 1$) and low π_c. All six prior choices have difficulties to identify varying levels of τ.

5.8 Discussion

Based on previous research, it is generally accepted that the choice of prior distribution on τ largely impacts the posterior interval estimation of δ in a meta-analysis of a few small trials [13, 131, 144, 145]. We demonstrated that in very sparse settings measures, such as the overall posterior median of δ, can become very inconsistent under alternative priors on τ as well. Even though, the final choice of prior should take into account the specific characteristics of each conducted meta-analysis, a solution in such sparse conditions would be to identify prior shapes that show robustness in the operational features of the posterior estimation of δ.

In this study we demonstrated that priors which place mass on small values of τ but sufficiently support larger value as well (Type A priors, eg. AU - $Uniform(-10, 10)$ prior on $log(\tau^2)$ scale) showed on average robust behaviour in most scenarios, followed by the $DuMouchel$ empirical prior (E), in comparison to other choices. Type D priors such as the dn - $Half\text{-}normal(0, 1)$ on τ, a prior that has been compared under an approximate normal setting and has been evaluated in settings of a few small trials [13, 147], did not perform satisfactorily neither under large levels of true heterogeneity nor under different settings of trial size and number of trials. Type A priors and $DuMouchel$ empirical prior place larger uncertainty around τ (Table 5.2) and produce a more data-driven inference on δ, in comparison to Type D priors such as the $Half\text{-}Normal$ (dn) prior or the more informative $Uniform(-10, 1.386)$ on $log(\tau^2)$ scale prior (du), which produces a more prior-driven inference on δ. Furthermore, we demonstrated that the use of priors with either a less restrictive or very confining prior range may be equally problematic, in terms of operational features and robustness.

Findings in perspective

Our study extends previous research on Bayesian hierarchical models' evaluations [131, 144, 145] in sparse-events MA of small populations. Contrary to previous evaluations on priors for heterogeneity [13, 144, 145, 147], we focused on a sparse-event setting, we then grouped the evaluated priors based on their shape. Except for observing the expected variations in the posterior intervals of δ, we observed a variation in the posterior medians of δ as well. Namely, priors that favour small τ are the ones that misestimate δ the least at very low event rates.

We further noticed a general overestimation when δ is large, as well as to a smaller extent

when δ takes smaller values. The primary reason for the overestimation of δ is the nature of a dichotomous outcome. For positive δ, more events are observed in the treatment arm, especially when δ is large [241]. Events in the treatment arm combined with zero events in the control arm result in overestimation. We also applied an alternative model that applies larger variance to $logit(\pi_T)$ than to $logit(\pi_C)$ in comparison to model 5.1 ([144] and Model 2 in [234]). Conclusions remained comparable, though when the alternative model was applied an underestimation of δ was observed when π_c was very low.

The variance within a single study relative to the estimated heterogeneity between studies determines this study's impact on the overall inference for δ. Naturally, small studies with zero events would produce a large within-study variability (standard errors) around the logOR study-specific effect which decreases the study's impact on the posterior overall effect. However, prior distributions that favour large values for τ allow small studies to have a larger weight. As a result, the contribution of small studies with one or two reported zero arms in a MA is enhanced when considering priors that support large τ. In both examples we reviewed herein, the increasing weight of studies with no observed events, mostly in a single arm, explains why the posterior median of δ are overestimated when less restrictive priors are applied (Figure 5.2 and Supplementary material I - Table 2). Therefore, in combination with the observed unstable study-specific treatment effect issues, alternative prior assumptions may enhance the impact of zero events in a few small trials MA, inducing a "small MA zero-event" bias on δ.

Main limitations

This work is subject to the assumption of normality for the study-specific effects and the overall treatment effect, by placing a weakly diffused normal prior on δ_i and δ; instead other dependence structures between δ and δ_i may be preferred [174]. Despite its common use, this assumption may not be appropriate considering the small number of studies and sparsity of events. Model 5.1 further assumes that the $logit(\pi_{iT})$ and $logit(\pi_{iC})$ have equal variances. This can be a restrictive assumption for which alternatives have been discussed [244]. In addition, other priors forms on μ_i, δ_i or δ may be considered such as: a $Uniform$, a $Student\text{-}t$, a $Truncated\text{-}t$ or a $Cauchy$ prior [118]. After partially evaluating these options through simulation, we did not observe changes in our conclusions. In the setting of a few small trials,

informative empirical priors that are based on published MAs of the Cochrane database can be used in a new MA of binary outcomes [152, 154]. However, such empirical priors have not been yet tailored for meta-analyses in rare diseases and therefore, may not be representative of heterogeneity commonly observed in such cases [152, 153, 199]. Based on preliminary non-reported results, such priors are expected to result in suboptimal frequentist characteristics similar to the very informative priors studied herein. Another restrictive option, given the small sample sizes, would be to model the studies as covariates and avoid the normal random-effects assumption.

In the simulation study we focused on positive treatment effects with low control event rates but not negative treatment effects with larger control event rates assuming that such effects are symmetric and their probabilities of success are reversed between the treatment arms.

The behaviour of a Bayesian MA might depend on the type of binary effect measure (log odds ratio, log risk ratio, risk difference). Such alternative measures could be of importance with sparse events MAs when normal approximations do not hold or when the logOR is undefined [149, 214].

Finally, one should consider the issue of inefficient Markov chain Monte Carlo sampling for rare events [245, 246]. In such extremely sparse settings, our findings might be sensitive to the sampling engine of the simulation study. Regardless of the sampler applied, we recommend conducting a formal convergence analysis in such sparse settings.

5.8 Conclusion

To conclude, a random-effects MA using a Bayesian binomial-normal hierarchical model has the potential to deal with high levels of zero events. The sensitivity of Bayesian models to the choice of priors is confirmed and produces not only diverse credible intervals but also diverse posterior medians for the overall treatment effect (δ). We showed that when performing a Bayesian binomial-normal MA under such sparse conditions, robust priors should have more mass close to zero, while supporting very large values as well (i.e. a less informative $Uniform(-10, 10)$ prior on $log(\tau^2)$). Priors that support only large or only mainly small values of heterogeneity (τ) result in substantial misestimation of δ in such sparse settings and

should be avoided. Except for robustness researchers should aim to account for the specific characteristics of each conducted meta-analysis before choosing a prior and setting prior levels of expected heterogeneity.

Supplementary material can be found at https://doi.org/10.1002/pst.20539.

Chapter 6

Borrowing strength from early phase outcomes in orphan drug development

K Pateras
S Nikolakopoulos
KCB Roes

Under revision

Chapter 6

Abstract

In drug development programs, proof-of-concept Phase II clinical trials typically have a biomarker as a primary outcome, or an outcome that can be observed with relatively short follow-up. Subsequently, the Phase III clinical trials aim to demonstrate the treatment effect based on a clinical outcome, that often needs a longer follow-up to be assessed. Short-term outcomes or biomarkers are typically associated with long-term outcomes and they are often included in Phase III trials. The decision to proceed to Phase III development is based on analysis of the short term outcome data from Phase II. In rare diseases, it is likely that only one Phase II trial and one Phase III trial are available. Positive results of the short term outcome Phase II trial are then likely seen as supporting (or even replicating) positive Phase III results on the long term outcome, without formal assessment and without accounting for between-study variability. We used double regression modelling applied to the Phase II and Phase III results to numerically mimic this informal assessment. We provide an analytical solution for the bias and mean square error of the overall effect that leads to a corrected *double-regression*. We further introduce a flexible Bayesian *double-regression* approach that also accounts for additional variance between the Phase II and Phase III trials. Such an approach includes the Phase II short term outcomes in the overall effect estimate of the primary outcome weighted by the extent to which they are in line with the Phase III short-term outcome results. We illustrate all methods with an orphan drug example for Fabry disease.

6.1 Introduction

Drug development programs typically include exploratory (Phase II) and confirmatory (Phase III) Randomized Controlled Trials (RCTs) to assess the efficacy, safety and appropriate dosages of an experimental (new) treatment. For regular "large disease" drug development programs decisions to conduct a Phase III trial are based on positive Phase II trials. If evaluated together they may induce a form of sampling-based selection bias (the succeeding trials are only conducted when the first trials were positive). Such a bias is usually not an issue, as more than one pivotal Phase III RCT will be required, such that the confirmatory evidence from Phase III can stand fully on its own.

Commonly in rare diseases, no more than two RCTs are conducted, one exploratory and one confirmatory [100]. The duration of the first exploratory trial is usually shorter than the duration of the succeeding confirmatory trial [247], hence Phase II primary endpoints are biomarkers or short-term clinical outcomes. Phase III primary clinical outcomes are often observed after a considerable time (long-term outcomes), therefore, even if $N = N_1 + N_2$ number of patients participate in both trials, only N_2 patients will be observed long enough to provide responses for the primary clinical outcome of interest. Biomarkers and secondary clinical outcomes are often observed earlier (short-term outcomes) and, therefore, easily included in both trials and, hence, available for all N patients. After both trials have been conducted, inference on the treatment efficacy is typically performed by evaluating the long-term responses of N_2 patients. In a rare disease setting, N_2 may not be large enough to solidly confirm treatment efficacy. In assessing the totality of evidence, the positive results from the Phase II trial could be then seen as supportive, as typically the short term outcome would be assumed to be associated with the long term outcome.

For example, Galafold (migalastat) acquired marketing authorization as an orphan drug for the treatment of Fabry disease in 2016 within Europe. Fabry disease is a rare, progressive disorder with an estimated prevalence of 1:117,000 to 1:40,000 [248]. The condition affect major organs and may result in life-threatening events. Until then, standard treatment for Fabry disease consisted of Enzyme Replacement Therapy [248]. Two main studies were submitted during the marketing authorization of migalastat; one randomized, placebo-controlled (AT1001-011, migalastat vs. Placebo) superiority study and one active comparison

randomized trial (AT1001-012, migalastat vs. Enzyme Replacement Therapy), with a non-inferiority design.

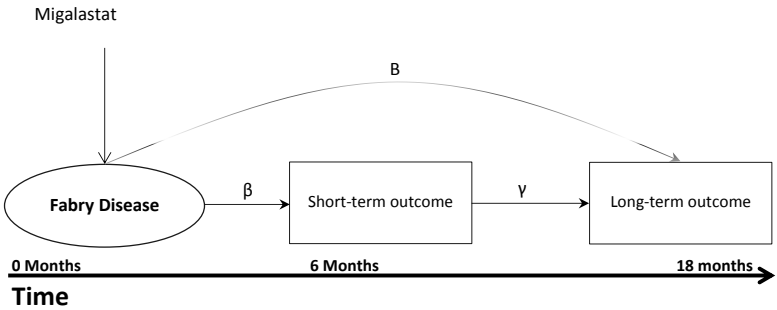

Figure 6.1: Relation between treatment vs. short-term outcome, treatment vs. long-term outcome and short-term vs. long-term outcome in Fabry disease example.

In trial 011 patients switched to migalastat 6 months post-randomization, while in trial 012 primary follow-up was considerably longer, with switching taking place 18 months post-randomization. In the first trial, the change in average globotriaosylceramide (GL-3) inclusions from baseline to six months was the primary outcome which produced non-conclusive evidence. The second trial utilized the annualized change in glomerular filtration rate ($eGFR$) at month 18 as the primary clinical outcome (Table 6.1). Both GL-3 and the annualized change in $eGFR$ at month 6 were collected in both trials. No strong correlation has been established in the literature between the GL-3 outcome and the change in glomerular filtration rate ($eGFR$) [249]. In study 011 after 6 months of treatment with migalastat 150 mg, eGFR values increased, whereas in the placebo treated group eGFR values declined [248]. This outcome among other secondary results led to the conduct of study 012. In trial 011, all patients treatment switched to migalastat at 6 months, an action that restricts the observation of a treatment effect on the primary long-term outcome. Given the limited available data, evidence from both trials were used for the final approval decision.

Table 6.1: Main randomized studies described in the European Public Assessment Report of Galafold

Study Number	Duration	Annualised rates of change in $eGFR$ from baseline to month 6	Annualised rates of change in $eGFR$ from baseline to month 18	Sample size	Start date
AT1001-011	6 months	Collected	Not Collected	67	Aug 2009
AT1001-012	18 months	Collected	Collected	52	Dec 2010

Analysis methods that use the relation between short and long-term outcomes may be applied to formally synthesize the evidence on treatment efficacy across the two trials. Engel and Walstra [250] formulated a *double-regression* (DR) approach, which can aid in more precise treatment effect estimation, by accounting for unobserved long-term outcome responses via observed short-term outcome responses. Their method utilizes the correlation to inform the point and variance estimates of the treatment effect on the long-term outcomes. For large samples their method has the proven potential to increase precision. However, for small sample sizes this may not be necessarily true [251]. Previously, in RCTs the *double-regression* approaches have been suggested mainly to inform treatment selection during interim analysis in seamless Phase II/III designs [252, 253, 254]. A Bayesian *double-regression* (BDR) analogue can be readily constructed [255] which maintains similar limitations to the frequentist alternative but could flexibly model the two Phase III outcomes' data and it can include additional historical trial data (i.e. Phase II short-term outcome data) as a prior distribution [256].

In this article we investigate how to model and estimate the efficacy of a new treatment on the long-term clinical outcome, using the data on short-term outcomes from both trials. We investigate methods that either account or do not account for the potential sampling-based selection bias when combining the Phase II and Phase III trials. We investigate a bias corrected *double-regression* approach and a flexible Bayesian approach regarding their performance to estimate the treatment effect on the long-term outcomes.

We focus on two related key problems: (1) the magnitude of the type 1 error inflation and bias when combining results from Phase II and III and (2) how to estimate the treatment effect on the long term outcome, using results from both studies and assess this estimate in terms of bias and variance.

The paper is organized as follows. First, we formalize the problem with a bivariate linear

model, then we introduce its conditional form and briefly discuss specific model variations, e.g., the *single-regression* approach. We provide an approximate analytical solution to the problem of sampling-based selection bias moving from Phase II to Phase III based on the Phase II short-term outcome in section 6.3. In section 6.4, we propose a Bayesian solution to the estimation problem, a model that down-weights the impact of short-term data via a historical power prior. This prior dynamically accounts for the bias in estimating the same treatment effect across the two trials, while accounting for additional between-trial variance (heterogeneity - τ) around the short-term outcome effect. Finally, we illustrate the applicability of methods, using a simulation study, in section 6.5. The article ends with a discussion and steps for further research.

6.2 Models for the joint Phase II and III data

Consider a Phase II trial of total sample size N_1 and a Phase III trial of total sample size N_2. For both trials it is assumed that equal number of patients ($n_k = N_k/2,\ k = 1, 2$) are randomized to the control and experimental treatment. Let us denote Y_k the long-term treatment response for patients in trial k and X_k the short-term treatment response in trial k, $k = 1, 2$. For the remainder of the manuscript we use bold letters to denote patients' dimension vectors $\mathbf{Y} = \{Y_i\}$ where $i = 1, ...N_k$. We denote as $\mathbf{Y}$ and $\mathbf{X}$ the long-term or short-term outcome data which correspond to patients of both Phase II and Phase III trials. Of these, $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{Y}_2$ are observed, while $\mathbf{Y}_1$ is not observed.

Bivariate modelling for short-term and long-term outcomes between studies

The long-term and short-term outcomes are assumed to follow a bivariate normal distribution and are modelled as $(\mathbf{X}, \mathbf{Y})$,

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N \left[\begin{pmatrix} \alpha + \beta \mathbf{t} \\ A + B\mathbf{t} \end{pmatrix}, \quad \sum \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right] \tag{6.1}$$

where σ_x^2 and σ_y^2 denote the true outcomes variances, ρ the true correlation between the two outcomes and $\mathbf{t}$ a vector indicating whether the i^{th} patient receives control or experimental

treatment. Throughout the manuscript we assume that the between-study correlation equals to zero (Online supplementary - A3).

The above bivariate model can be conditionally expressed as

$$\begin{aligned} \mathbf{X}|\mathbf{t} &\sim N(\alpha + \beta\mathbf{t}, \sigma_x^2) \\ \mathbf{Y}|\mathbf{t},\mathbf{x} &\sim N(\mathrm{a} + b\mathbf{t} + \gamma\mathbf{x}, \sigma_0^2) \end{aligned} \tag{6.2}$$

where $\sigma_0^2 = \sigma_y^2 - \gamma^2\sigma_x^2$, $\mathrm{a} = A - \gamma\alpha$, $b = B - \gamma\beta$ and $\gamma = \rho\sigma_y/\sigma_x$

Double-regression to estimate the effect of primary long-term outcome

At the end of both trials short-term outcome data $\mathbf{X}$ for $N = N_1 + N_2$ patients and long-term outcome data $\mathbf{Y}_2$ for only N_2 patients are observed. $\mathbf{Y}$ corresponds to the outcome of interest related to which estimation and hypothesis testing will be performed. The *double-regression* utilizes the relation between short-term and long-term outcomes and allows estimation of the long-term outcome parameter B.

Based on the *double-regression* method, parameters α, β and σ_x^2 are estimated via the regression of $\mathbf{X}|\mathbf{t}$ on N patients, as $\hat{\alpha}, \hat{\beta}$, s_x^2 and parameters a, b, γ and σ_0^2 are estimated via the regression of $\mathbf{Y_2}|\mathbf{X_2}, \mathbf{t}$ on N_2 patients, as $\hat{\mathrm{a}}, \hat{b}, \hat{\gamma}$, s_0^2, $s_y^2 = s_0^2 + \hat{\gamma}^2 s_x^2$, $\hat{A} = \hat{\mathrm{a}} + \hat{\gamma}\hat{\alpha}$, $\hat{\rho} = \hat{\gamma}s_x/s_y$ [250, 253]. The primary effect of interest B is then estimated via:

$$B = b + \gamma\beta \tag{eq1}$$

The variance of $\hat{B}$ is shown in [250] to be equal to

$$var(\hat{b}) + \gamma^2\ var(\hat{\beta}) + \beta\ var(\hat{\gamma}) + 2\beta\ cov(\hat{b}, \hat{\gamma})$$

estimates of the above can be obtained by using the individual estimates acquired from the regression analyses (model 6.2). Under model 6.2, hypothesis testing is performed as $H_0 : \hat{B} \leq 0$ vs. $H_1 : \hat{B} > 0$ via $Z^y_{1-\alpha} < \hat{B}/\sqrt{var(\hat{B})}$,where $Z^y_{1-\alpha}$ is the $(1-\alpha)^{th}$ standard normal critical quantile. A direct Bayesian analogue to the conditional model 6.2 has been discussed elsewhere [255]. This Bayesian model has been shown to produce comparable results to model 6.2 under diffuse "non-informative" priors for each parameter in general settings.

Flexible Bayesian (double-) regression

We can model the Phase II short-term outcome data ($\mathbf{X}_1$) via a Bayesian *single-regression* and we can utilize the posterior distribution as prior on a Bayesian *double-regression* model on the Phase III short-term outcome data as follows. Let us assume a bivariate variable $(\mathbf{X}_2, \mathbf{Y}_2)$ of dimensions $N_2 \cdot 2$ with a covariance matrix Σ_2. Barnard et al suggested decomposing Σ_2 and applying independent priors on ρ, σ_{x_2} and σ_{y_2} [255, 257]. In our two-dimensional scenario, a multivariate normal likelihood could be specified on the short-term and long-term Phase III outcome data by conditional distributions as follows

$$\begin{aligned} \mathbf{X_2}|\mathbf{t} &\sim N(\alpha + \beta\mathbf{t}, \sigma^2_{x_2}) \\ \alpha &\sim N(\mu_\alpha, \sigma^2_\alpha), \beta \sim N(\mu_\beta, \sigma^2_\beta) \\ \mathbf{Y_2}|\mathbf{t}, \mathbf{x}_2 &\sim N(A + B\mathbf{t} + \rho\frac{\sigma_{y_2}}{\sigma_{x_2}}(\mathbf{x_2} - \mu_{x_2}), (1-\rho^2)\sigma^2_{y_2}) \\ a &\sim N(0, 10^2), b \sim N(0, 10^2) \end{aligned} \tag{6.3}$$

where μ_{x_2} denotes the mean value of the short-term Phase III outcome data. A prior has to be placed on ρ parameter, e.g. prior $\rho \sim Uniform(-1, 1)$ uniformly weights our prior considerations around the correlation parameter. We can further assume simply two half-normal priors on $\sigma_{x_2}, \sigma_{y_2} \sim HN(0, 1)$. To mimic model 6.2 we inform the σ_{x_2} prior based on the posterior model variance from Phase II short-term outcome data i.e. fitting them over an optimized log-normal prior distribution. In order to further mimic model 6.2 we

have set normal distribution priors based on Phase II posterior effect and variance mean estimates of the short-term outcome parameters ($\mu_\alpha, \mu_\beta, \sigma^2_\alpha, \sigma^2_\beta$). In comparison to the direct Bayesian analogue of model 6.2, where the strength of the relationship between short and long-term endpoints becomes clear only after combining the posterior mean estimates via the γ parameter, model 6.3 is more intuitive, as it directly models the correlation (ρ) between the two outcomes and it directly produces posterior Markov Chain Monte Carlo draws from B. Therefore, under such a fully Bayesian approach there is no need for numerical addition of treatment effect mean estimates. Posterior inference can be obtained via traditional Markov Chain Monte Carlo application software (i.e. JAGS [185]) or even analytically under convenient prior distributions [256]. In this Bayesian model we assume that hypothesis testing for H_0 vs H_1 will be performed by utilizing posterior probabilities as $Pr(B > 0|\mathbf{Y}) > \omega$ where $\omega = 0.95$.

If we set the correlation very close to zero; i.e. $\rho \sim U(-0.01, 0.01)$, then, the Phase III trial long-term outcome data are evaluated individually under a standard (Bayesian) linear *single-regression* model. In comparison to the *single-regression* models, the advantage of models 6.2, Bayesian 6.2 and 6.3 rest in their ability to numerically calculate the impact of accounting for the Phase II short-term outcome data in analysing the long-term outcome. Additional details of the *(Bayesian) single-regression* models can be found in the online supplementary (A1).

6.3 Type 1 error inflation and bias due to selection based on short term outcome results

Usually, a Phase II decision leads to the initiation of a Phase III trial. This decision can be based on a test statistic for the early short-term outcome and an imposed critical value; i.e. $z_{1-\alpha}$. This is clearly an oversimplification of the actual Phase II to Phase III transition decision, but used here to illustrate the potential impact on type 1 error and bias if the results are combined. In this simplified model, the distribution of the available Phase II trial short-term outcome $f(\mathbf{X}|Z_{X_1} > z_{1-\alpha})$, will be truncated, where Z_{X_1} denotes the standardized difference of the short-term Phase II trial outcome. If the analysis of Phase III data occurs independently from earlier Phase trial data, we expect no increase of Type I error and bias, though the power might remain low due to the limited trial sample size. In the assessment of totality of evidence

in this rare disease setting, however, positive results from both the Phase II trial and Phase III trial may well be seen as reinforcing. This informally combines evidence between trials which often results in positively biased inferences in favour of the long-term treatment effect B, while an error inflation is observed in the double-regression long-term outcome inference (models 6.2, 6.3 and Figure 6.2). In such situations, the bias on $\hat{B}$ estimate, based on model 6.2 is given by the following approximation (Appendix of Chapter 6 - A2),

$$Bias(\hat{B}) = \sigma_y^{'} \frac{w_1 \rho \lambda \sigma_{x_1}}{\sigma_x^{'} \sqrt{N_1/2}} \tag{eq2}$$

where $\sigma_y^{'2} = \sigma_y^2 + \gamma^2 D$, $\sigma_x^{'2} = \sigma_x^2 + D$, $\lambda = \frac{\phi(\omega)}{1 - \Phi(\omega)}$, $\omega = \frac{Z_{1-\alpha} - \mu_{x_1}}{\sigma_{x_1}/\sqrt{N_1/2}}$, $w_1 = N_1/N$. ϕ and Φ are probability density and cumulative functions of the standard normal distribution respectively, $D = w_1\Big((2\sigma_1^2/N_1)\zeta + A^2(1 - w_1^2 - w_2^2) + 2A(\mu_{x_1} - \mu_x)\Big)$, $A = (\sigma_1/\sqrt{N_1/2})\lambda$, $\zeta = a\lambda - (\lambda)^2$ and N_1 denotes the sample size of the Phase II trial.

An approximate value for $MSE(\hat{B})$ of the double-regression is equal to (Appendix of Chapter 6 - A2)

$$MSE(\hat{B}) = \underbrace{2\sigma_y^{'2}\Big(\frac{w_1 \rho \lambda \sigma_{x1}}{\sigma_x^{'}\sqrt{N_1}}\Big)^2}_{Bias(\hat{B})^2} + \underbrace{2\sigma_y^{'2}\Big(\frac{1-\rho^2}{N_2} + \frac{\rho^2}{N}\Big)}_{Var(\hat{B})} \tag{eq3}$$

As we observe in eq3, the inflation in MSE depends on (i) the decision threshold to initiate the Phase III trial through λ parameter, (ii) the Phase II short-term outcome mean (μ_{x_1}) and variance $(\sigma_{x_1}/\sqrt{N_1/2})^2$, (iii) the number of patients in the Phase II trial (N_1) and (iv) and the magnitude of the correlation (ρ). An increase in σ_{x_1} results in an increase of MSE, while as N_1 decreases, the MSE increases as well. A similar behaviour is observed in terms of Type I error (Figure 6.2). More specifically, Type I error rates increase considerably with higher ρ, while the power curves, in general, increase with more patients being allocated to the Phase III trial (N_2) (Figure 6.2).

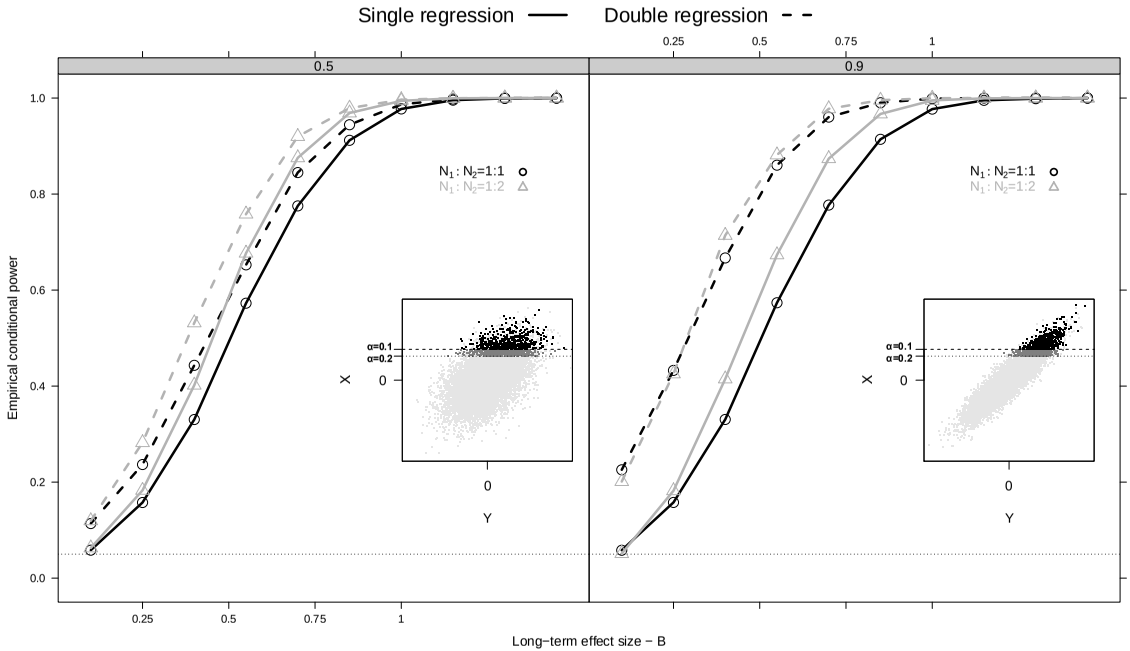

Figure 6.2: Conditional power curves comparing the performance of the single and double-regression for the following scenarios; $N_1{:}N_2 \in \{1{:}1, 1{:}2\}$, $\beta_1 = 0$, $\sigma_y^2 = \sigma_x^2 = 1$, $\rho_r \in \{0.1, 0.9\}$, $N = 120$, $\alpha_x = 0.1$ and $B, \beta \in \{0, 0.1, 0.2, ..., 1\}$. No additional between-trial variation (τ) was introduced in this set up and each scenario was replicated 10,000 times. The inner figures serve as an explanation to the observed type I error increase, as they present the joint strict null hypothesis ($B = \beta = 0$) distribution of the short and long-term treatment effect for the Phase III trials (light gray dots) and the truncated, based on a positive decision criteria, Phase II trials (black and dark grey dots). When utilizing the Phase II trials (darker dots in the inner Figures), larger critical levels result in an average overestimation of the treatment effect which consistently produces an average increase in error rates and on average larger bias is incorporated in the final inference. This mean increase can be observed in the expression of mean square error for the long-term treatment effect estimate (eq3). As expected based on eq3, all error rates increase with higher ρ and the power curve increases with lower σ. A similar behaviour was observed between the equivalent Bayesian single-regression and Bayesian double-regression alternative.

Based on the aforementioned bias and mean square error expressions and by replacing parameters with their estimates, the long-term outcome effect and variance of a (bias) corrected *double-regression* model are estimated as (Appendix of Chapter 6 - A2),

$$\begin{aligned} \hat{B}^{'} &= \hat{B} - B\tilde{ias}(\hat{B}) \\ v\tilde{ar}(\hat{B})^{'} &= 2(s_y^2 - \hat{\gamma}^2\hat{D})\Big(\frac{1-\hat{\rho}^2}{N_2} + \frac{\hat{\rho}^2}{N}\Big) \end{aligned} \tag{6.4}$$

6.4 Effect of additional short-term outcome heterogeneity (τ)

All above models assume that the between-study variability of the short and long-term outcomes equal to zero ($\tau_x = \tau_y = 0$) and therefore, all N observations are derived from the same population. Phase II vs. Phase III trials typically do not have similar protocols, as the Phase II trials are usually more restrictive in patient inclusions, therefore, exploring additional between-study variance becomes important, especially in a rare disease setting where each disease contains highly heterogeneous patient populations. Such between-study variability is expected to influence the short-term and long-term treatment effect estimates. A proper estimation of τ with just two available studies is not likely reliable [11, 178, 241].

We utilize a mechanism based on power priors to account for the between-study variability within a Bayesian framework [258]. By estimating a power parameter η that represents the conflict between the short-term outcome data of the two available trials, model 6.3 can be further extended to account for the short-term outcome excess between-study variance (τ), along with the bias [258, 259, 260].

Bayesian flexible double-regression

Let us assume that data $\mathbf{X}_1$ exist for the short-term outcome from the Phase II study and $\mathfrak{B}$ are a set of linear regression parameters. Given the definition of a power prior [261], the posterior distribution after observing the Phase II short-term outcome data would be

$$\pi(\mathfrak{B}|\mathbf{X}_1, \eta) \propto L(\mathfrak{B}|\mathbf{X}_1)^{\eta}\pi_0(\mathfrak{B})$$

Then, the posterior for $\mathfrak{B}$ after observing the Phase III study's short-term outcome data ($\mathbf{X}_2$) would be

$$\pi(\mathfrak{B}|\mathbf{X}, \eta) \propto L(\mathfrak{B}|\mathbf{X}_2)L(\mathfrak{B}|\mathbf{X}_1)^{\eta}\pi_1(\mathfrak{B})$$

The posterior distribution of $\mathfrak{B}|\mathbf{X}_1$ in the normal case ([262]) is know to be equal to

$$\mathfrak{B}|\mathbf{X}_1, \eta \sim N\Big((\mathbf{T}_1^{'}\mathbf{T}_1)^{-1}\mathbf{T}_1^{'}Y_1, \frac{\sigma_{x1}^2}{\eta}(\mathbf{T}_1^{'}\mathbf{T}_1)^{-1}\Big) \quad \text{(eq4)}$$

, where $\mathbf{T}_1$ is the design matrix with column vectors $\mathbf{1}$, $\mathbf{t}$. We consider the following conditional model

$$\begin{aligned}
\mathbf{X_2}|\mathbf{t} &\sim N(\alpha + \beta\mathbf{t}, \sigma_{x2}^2) \\
\alpha &\sim N(\mu_\alpha, \sigma_\alpha^2/\hat{\eta}), \beta \sim N(\mu_\beta, \sigma_\beta^2/\hat{\eta}) \\
\mathbf{Y_2}|\mathbf{t}, \mathbf{x}_2 &\sim N(A + B\mathbf{t} + \rho\frac{\sigma_{y2}}{\sigma_{x_2}}(\mathbf{x_2} - \mu_{x_2}), (1-\rho^2)\sigma_{y2}^2) \\
A &\sim N(0, 10^2), B \sim N(0, 10^2)
\end{aligned} \quad (6.5)$$

The conditional set-up of model 6.5 remains similar to 6.3. Now dynamic informative power priors based on eq4 are placed on the short-term endpoint's parameters α and β. Such priors control the borrowing of the historical data and discount the short-term prior in case of treatment effect's disagreement.

Table 6.2: Summary of aforementioned statistical methods

Abbreviation	Model	(F)requentist/(B)ayesian?	Long-term/Short-term?	Phase (II/III)?	Heterogeneity?
(B)SR	(Bayesian) single-regression	F/B	Long	III	No
(B)DR	(Bayesian) double-regression	F/B	Short&Long	II+III	No
DRC	Double-regression corrected	F	Short&Long	II+III	No
BFDR	Bayesian flexible double-regression	B	Short&Long	II+III	Indirectly

Estimation of η

A number of power prior (guided-value) formulations have been suggested [258, 259, 260]. Among the above alternatives, we chose one that selects a guided-value that maximizes the marginal likelihood [260]. The guide value of η based on the marginal likelihood criterion has an estimate of

$$\hat{\eta} = \underset{0<\eta\leq 1}{argmin}[-2log\{m(\eta)\}] \quad \text{(eq5)}$$

where $m(\eta)$ is the marginal likelihood. Ibrahim et al [262] provided an analytical expression of $-2log\{m(\eta)\}$ for the normal outcome case. Figure 1 in the online supplementary (A4) presents the empirically calculated relationship between η and varying levels of β_1.

In model 6.5, similarly to model 6.3, we are interested in the overall long-term outcome effect B and we assume that hypothesis testing for H_0 vs H_1 will be performed by utilizing posterior probabilities as $Pr(B > 0|Y) > \omega$ where $\omega = 0.95$.

6.5 Simulation study

A simulation study was conducted to assess the relative performance of the suggested methods, in the analysis of the Phase III long-term outcome data. For illustrative purposes, we assume that the two available Phase II and Phase III trials had a similar control treatment, therefore, the second trial would have been designed as a placebo-controlled trial. In this section, we assume that the decision to conduct the second Phase III trial was taken on the basis of available evidence in the first Phase II trial on a single short-term outcome. At the end of the second trial, individual data of N patients are available on the short-term and data of N_2 are available on the long-term outcomes. The simulation study results were derived from a bivariate normal model simulation strategy as described in the online supplementary (A3).

The *single-regression* (SR), *double-regression* (DR), corrected *double-regression* (DRC) methods ignore τ_x and therefore assume a different underlying data generating model in comparison

to the Bayesian flexible *double-regression* (BFDR) approach. Even though, they are not directly comparable (Table 6.2), we empirically compared the four aforementioned statistical methods by generating at least 10,000 simulated combinations of the two available trials data. To do so, we simulated scenarios of the final trial analysis on the primary long-term endpoint assuming a variety of combinations between the short (β) and long-term (B) outcome treatment effects. The latter were varied as (Scenario I) $B = \beta = 0$, (Scenario II) $B = \beta = 0.6$ (Scenario III) $B_1 = 0, B_2 = 0.2, \beta_1 = 0, \beta_2 = 0.2$ and (Scenario IV) $B = 0.6, \beta = 0$, we assumed that $\rho = 0.9$, $\alpha_x \in \{0.05, 0.1, 0.2\}$, the between-study standard deviation equal to $\tau = 0$, while all variances were set equal to 1. Specific alternative versions of scenarios I and II were produced by varying ρ and τ. The first (I) scenario describes variations of the strict null ($\tau = 0$) and null hypothesis under heterogeneity ($\tau = 1$), while the second (II) scenario describes a common alternative hypothesis on both outcomes and trials. Scenario III can occur when heterogeneous populations are selected for the Phase II and Phase III trial, while the fourth (IV) scenario describes a situation where the long-term outcome true effect exists but the short-term outcome equals to 0. All remaining settings (ie. number of trials (k), total sample sizes N, sample size ratio between trials $N_1 : N_2$) were reflective of a typical rare disease setting and based on the Galafold example (Table 6.1). All simulations were performed via R [263] and JAGS [185].

6.6 Results

(Strict) null hypothesis scenario (I: $B = \beta = 0$)

The Bayesian flexible *double-regression* (BFDR) results in treatment effects closer to the *single-regression* (SR) estimates than the *double-regression* (DR) approach under the null hypothesis simulation (Scenario I - Table 6.3). The corrected *double-regression* (DRC) approach presents a similar behaviour producing long-term effect estimates even closer to the SR than the BFDR approach. In the three null hypothesis scenarios I(b-d) ($B = \beta = 0$), DR results in the largest estimated treatment effect and produces the largest type I error inflation while DRC generally inflates the Type I error the least among the three investigated methods. An interesting exception, that we further discuss in section 6.7, is observed in scenario Ia, where the BFDR approach produces stricter error rates than the DRC approach. In general, the SR method controls type I error the most, while the DR method controls type I error the least.

Table 6.3: Long-term conditional average treatment effect estimates (means, posterior means, confidence intervals, credible intervals) and average treatment efficacy p-values and probabilities of the four models (Table 6.2) given that $\rho = 0.9$, $\tau = 0.01$, $\rho_B = 0$ and σs=1, except where noted otherwise, based on at least 10.000 simulations.The first line SR of each scenario (I) presents a frequentist *single-regression* on the Phase III long-term outcome data. DR correspond to the frequentist *double-regression*. Last, the DRC lines present the result for the bias corrected *double-regression* approach and the BFDR lines present the results for the Bayesian flexible *double-regression* approach.

Scenario	Model	Mean/Posterior mean B	Error/Power
	α_x:	$(0.05 \cdot 0.1 \cdot 0.2)$	$(0.05 \cdot 0.1 \cdot 0.2)$
Ia. $B = \beta = 0$	SR	0.001 · 0.003 · 0.002	0.057 · 0.054 · 0.053
	DR	0.256 · 0.220 · 0.178	0.318 · 0.247 · 0.183
	DRC	0.087 · 0.075 · 0.063	0.079 · 0.066 · 0.060
	BFDR	0.170 · 0.156 · 0.133	0.054 · 0.037 · 0.022
b. $B = \beta = 0$ $\rho = 0.5$	SR	0.000 · 0.003 · 0.002	0.055 · 0.053 · 0.054
	DR	0.141 · 0.123 · 0.100	0.148 · 0.130 · 0.114
	DRC	-0.028 · -0.022 · -0.015	0.045 · 0.047 · 0.048
	BFDR	0.089 · 0.083 · 0.071	0.070 · 0.066 · 0.056
c. $B = \beta = 0$ $\tau = 0.5$	SR	0.001 · 0.004 · 0.003	0.057 · 0.056 · 0.054
	DR	0.228 · 0.198 · 0.160	0.215 · 0.177 · 0.142
	DRC	0.059 · 0.053 · 0.045	0.073 · 0.069 · 0.066
	BFDR	0.150 · 0.141 · 0.120	0.102 · 0.088 · 0.069
d. $B = \beta = 0$ $\tau = 0.5$ $\rho = 0.5$	SR	0.000 · 0.003 · 0.002	0.054 · 0.054 · 0.054
	DR	0.126 · 0.110 · 0.090	0.127 · 0.113 · 0.104
	DRC	0.032 · 0.030 · 0.026	0.072 · 0.069 · 0.071
	BFDR	0.078 · 0.075 · 0.064	0.095 · 0.092 · 0.086

Alternative hypothesis scenario (II: $B = \beta = 0.6$)

In scenario II ($B = \beta = 0.6$), all methods identified a treatment effect close to the true value (Table 6.4). The empirical power to identify a treatment effect is usually large for the BFDR, and considerably larger for the DRC than SR approach. Among the DRC and BFDR methods, BFDR produces treatment effect means closest to the true value. In scenario IIa ($\tau = 0$), DRC performs better in terms of 95% coverage whereas in scenario IIb where $\tau = 0.5$, BFDR results in coverage closest to 95%.

Table 6.4: Long-term conditional average treatment effect estimates (means, posterior means, confidence intervals, credible intervals) and average treatment efficacy p-values and probabilities of the four models (Table 6.2) given that $\rho = 0.9$, $\tau = 0.01$, $\rho_B = 0$ and σs=1, except where noted otherwise, based on at least 10.000 simulations.The first line SR of each scenario (II,III,IV) presents a frequentist *single-regression* on the Phase III long-term outcome data. DR correspond to the frequentist *double-regression*. Last, the DRC lines present the result for the bias corrected *double-regression* approach and the BFDR lines present the results for the Bayesian flexible *double-regression* approach. In Scenario III the correction for the DRC method is calculated based on that true long-term effect is equal to 0.2.

Scenario	Model	Mean/Posterior mean B	Error/Power	95% coverage
	α_x:	$(0.05 \cdot 0.1 \cdot 0.2)$	$(0.05 \cdot 0.1 \cdot 0.2)$	$(0.05 \cdot 0.1 \cdot 0.2)$
IIa. $B = \beta = 0.6$	SR	$0.598 \cdot 0.596 \cdot 0.598$	$0.659 \cdot 0.655 \cdot 0.658$	$0.940 \cdot 0.940 \cdot 0.942$
	DR	$0.643 \cdot 0.625 \cdot 0.612$	$0.942 \cdot 0.924 \cdot 0.909$	$0.954 \cdot 0.952 \cdot 0.951$
	DRC	$0.634 \cdot 0.621 \cdot 0.611$	$0.935 \cdot 0.920 \cdot 0.907$	$0.956 \cdot 0.954 \cdot 0.952$
	BFDR	$0.632 \cdot 0.617 \cdot 0.607$	$0.663 \cdot 0.634 \cdot 0.612$	$0.997 \cdot 0.997 \cdot 0.997$
b. $B = \beta = 0.6$	SR	$0.600 \cdot 0.597 \cdot 0.598$	$0.581 \cdot 0.576 \cdot 0.576$	$0.899 \cdot 0.897 \cdot 0.897$
$\tau = 0.5$	DR	$0.654 \cdot 0.633 \cdot 0.617$	$0.803 \cdot 0.775 \cdot 0.753$	$0.910 \cdot 0.896 \cdot 0.884$
	DRC	$0.645 \cdot 0.629 \cdot 0.615$	$0.793 \cdot 0.771 \cdot 0.751$	$0.905 \cdot 0.893 \cdot 0.882$
	BFDR	$0.641 \cdot 0.624 \cdot 0.611$	$0.672 \cdot 0.647 \cdot 0.628$	$0.956 \cdot 0.948 \cdot 0.941$
III. $B_1 = 0, B_2 = 0.2,$	SR	$0.201 \cdot 0.205 \cdot 0.203$	$0.161 \cdot 0.159 \cdot 0.159$	$0.940 \cdot 0.943 \cdot 0.946$
$\beta_1 = 0, \beta_2 = 0.2$	DR	$0.357 \cdot 0.326 \cdot 0.288$	$0.382 \cdot 0.335 \cdot 0.286$	$0.912 \cdot 0.930 \cdot 0.943$
$\tau = 0.5$	DRC	$0.187 \cdot 0.182 \cdot 0.173$	$0.154 \cdot 0.143 \cdot 0.140$	$0.961 \cdot 0.963 \cdot 0.963$
	BFDR	$0.327 \cdot 0.308 \cdot 0.280$	$0.243 \cdot 0.215 \cdot 0.180$	$0.961 \cdot 0.969 \cdot 0.976$
IV. $B = 0.6, \beta = 0$	SR	$0.601 \cdot 0.601 \cdot 0.603$	$0.576 \cdot 0.576 \cdot 0.582$	$0.762 \cdot 0.762 \cdot 0.767$
$\tau = 0.5$	DR	$0.828 \cdot 0.828 \cdot 0.760$	$0.936 \cdot 0.936 \cdot 0.898$	$0.898 \cdot 0.898 \cdot 0.854$
	DRC	$0.659 \cdot 0.659 \cdot 0.645$	$0.817 \cdot 0.817 \cdot 0.799$	$0.748 \cdot 0.748 \cdot 0.732$
	BFDR	$0.750 \cdot 0.750 \cdot 0.719$	$0.764 \cdot 0.764 \cdot 0.758$	$0.904 \cdot 0.904 \cdot 0.900$

Scenarios III and IV

In scenario III ($B_1 = 0, B_2 = 0.2, \beta_1 = 0, \beta_2 = 0.2$), the BFDR produces similar findings to the DR approach, while the DRC method discards most Phase II information and its results are close to the SR approach (Table 6.4). DRC retains a comparable behaviour in scenario IV ($B = 0.6, \beta = 0$), where it discards most of the sample-based selection bias and it produces results closer to the analysis of the Phase III study alone. Though in the same scenario DRC produces very suboptimal 95% coverage in comparison to the other methods. In scenarios III, IV, as well as I, the naive pooling represented via the formal DR method, systematically and largely overstates our confidence in treatment efficacy.

6.7 Discussion

In a drug development procedure, it is not uncommon that positive Phase II results on short-term (biomarker) outcomes are not predictive of a Phase III success on long-term clinical outcomes. If Phase II and Phase III results are then assessed (perhaps informally) jointly to support efficacy, this assessment may be subject to selection bias and may increase uncertainty of the true treatment effect. Such an informal combination of results may increase to a great extent (more than 3 times) the Type I error rate of null hypothesis, rendering the combined true long-term treatment effect misleading. Especially in rare diseases, where the validation of short-term surrogate endpoints can become problematic, due to the small and often heterogeneous populations, the small sample sizes and the insufficient number of available trials, only long-term hard endpoints are usually appropriate to prove treatment efficacy.

In this manuscript, in addition to identifying and investigating the above issue, we explored methods that could be utilized in order for early and late Phase trial data to be combined while accounting for the underlying sampling-based selection bias. The flexible Bayesian *double-regression* includes the borrowing of historical information, while this model downgrades the historical prior upon short-term outcome data conflict. The corrected *double-regression* method approximately corrects the biased long-term mean effect and variance estimate.

In most explored scenarios, the corrected *double-regression* method inflates the Type I error and the incorporated bias the least in comparison to the *double-regression* and Bayesian flexible

double-regression methods. This behaviour is not observed in scenario Ia, where the Bayesian flexible *double-regression* controls better the Type I error. This possibly happens because the Bayesian flexible *double-regression* approach completely downgrades the impact of Phase II trial when its short-term treatment effect is different than the Phase III trial short-term treatment effect. Therefore, on average the Bayesian approach becomes less prone to false-positive results based on possible very positive Phase II short-term outcome trial effects when τ is low and/or ρ is high (see, black dots of inner right panel of Figure 6.2). On the contrary, the corrected *double-regression* method corrects the Phase II effects and then utilizes both Phase II and Phase III effects without heavily downgrading the Phase II results data upon data conflict.

Both the Bayesian flexible *double-regression* and the corrected *double-regression* methods would be an attractive solution to the increased Type I error of the informal combination of two small available trials. The consideration of these methods was shown to be rather important when, (i) the preceding Phase II trial conservatively resulted to the Phase III trial and/or (ii) the association of utilized short and long-term outcomes is high. An informal combination of results accross Phases often happens when both of the above hold, though, when neither holds then the complexity of suggested methods may outweigh the gains of their application.

Further optimized versions of the Bayesian flexible *double-regression* model could be developed and they may perform more optimally in comparison to the current. For instance, different and more optimal guided values could be applied on the flexible Bayesian *double-regression* [258, 259, 260]. All guided value formulations are expected to be somehow comparable and less of an issue as the power parameter is imposed on the short-term endpoint and indirectly affects the long-term primary endpoint. An alternative approach that controls type I error on the long-term outcome, while borrowing historical information, may also provide a more formal solution [259]. Future research could compare these alternatives vis-á-vis each other or with other methods. More covariates could be included, and then their performance could be tested with ease as all presented models are readily generalizable to full regressions. Lastly, in this work, we accounted for but did not estimate τ. Due to the only two available studies, a proper estimation of τ is currently known to be almost non-feasible [11, 150, 178, 241].

In the motivating example we assumed that both trials were superiority trials, while if

we had kept the initial designs, different strategies may have been more appropriate. Nonetheless, examples of two superiority trials, one Phase II and one Phase III, exist in the literature. For example, the drug development program of thalidomide for the treatment of multiple myeloma contained two randomized superiortiy clinical studies of similar design, a supportive (GISMM2001) and a main study (IFM 99-06), that compared melphalan-prednisone (control treatment) to thalidomide (experimental treatment) [247]. The supportive study was shorter and it reported clinical response rates and event free survival as primary endpoints. The main study was longer in duration and it reported overall survival, as main endpoint and clinical response rates and event free survival, as secondary endpoints. The suggested methodology could be tailored to account for the possibility of sampling-based selection bias under survival and other types of outcomes and even to combine different study designs.

Throughout the manuscript normality was assumed, an assumption that could be challenged with rare diseases sample sizes [100, 247, 248]. We approximated a truncated normal with a normal distribution with mean and variance equal to that of the former. This decision was made to aid calculations on the distribution mixture (Appendix of Chapter 6 - A2). Better approximations for the truncated normal distribution may exist, such as the chi-square distribution and their performance could be explored as well [264]. We should note that for moderately sized N_2 in comparison to N and small correlation between the two outcomes, a *single-regression* might be more efficient than a *double-regression*, due to the noise introduced by the short-term outcome [250].

In this article we performed a post-hoc combination of available information after the conduct of the Phase II and Phase III trial. However, it may be very relevant to (prospectively) plan to pool the data from both studies and to use the short term outcomes of the Phase II study to increase the precision, with which the efficacy on long term outcome is estimated overall [252, 253, 254]. An alternative strategy could be to conduct one single trial with interim analysis, then, based on the observed treatment effects on the short-term endpoints, to decide whether to proceed following up the patients [253].

To conclude, the often naive retrospective pooling of Phase II short-term outcome data to support the true long-term outcome data inference at the end of confirmatory Phase III trials

should be performed via formal numerical approaches. Such approaches should control the sampling-based selection bias, in order to avoid inflating the Type I error under the null hypothesis and prevent overestimating our beliefs on the treatment effect, especially in a small population context. We hope that this manuscript, except for introducing possible solutions, raises awareness of potential mishaps with ad-hoc combinations of trial outcome results.

Appendix - Derivation of $MSE(\hat{B})$ - (A2)

The $MSE(\hat{B})$ of the long-term outcome equals to

$$MSE(\hat{B}) = Bias(\hat{B})^2 + Var(\hat{B})$$

Appendix - Derivation of $Bias(\hat{B})$ (A2.1)

Let assume that $\sigma_{x_1}, \sigma_{x_2}$ are known for the Phase II and Phase III trials, then the short-term outcome treatment effect estimates are distributed as $\hat{\beta}_2 \sim N(\mu_{x_2}, \frac{2\sigma^2_{x_2}}{N_2^*})$ and $\hat{\beta}_1 \sim N(\mu_{x_1}, \frac{2\sigma^2_{x_1}}{N_1})$. In practice the Phase II short-term outcomes would follow an one-sided truncated normal distribution. The adjusted mean (μ_{x_1}) and variance ($\sigma^2_{x_1}$) of this short-term outcome one-sided truncated normal distribution $\hat{\beta}_1 \sim N_\alpha(\mu^{'}_{x_1}, \frac{2\sigma^{'2}_{x_1}}{N_1})$ equal to

$$\mu^{'}_{x_1} = \mu_{x_1} + \frac{\sigma_{x_1}}{\sqrt{N_1/2}}\lambda \quad \text{(eq3)}$$

$$\sigma^{'2}_{x_1} = \sigma^2_{x_1}\Big[1 + \zeta\Big] \quad \text{(eq4)}$$

where $\lambda = \frac{\phi(\omega)}{1 - \Phi(\omega)}$, $\zeta = a\lambda - \big(\lambda\big)^2$ and $\omega = \frac{Z^x_{1-\alpha} - \mu^{'}_1}{\sigma^{'}_1/\sqrt{N_1/2}}$ and ϕ and Φ are the probability density and the cumulative function of the standard normal distribution.

We assume that we can approximate a truncated normal with a normal distribution with updated mean and variance as follows $\hat{\beta}_1 \overset{approx}{\sim} N(\mu^{'}_{x_1}, \frac{2\sigma^{'2}_{x_1}}{N_1})$ ([265]). The overall $\hat{\beta}$ would be a mixture of the above density functions.

Given the set of two densities and weights (w_1 and w_2), such that $w_i \leq 0$ and $\sum w_i = 1$ the mixture can be represented as

$$f(x) = \sum_{k=1}^{2} w_k p_k(x)$$

The mean and variance of the above normal mixture of two distributions equal to $\mu_x = \sum_{k=1}^{2} w_k \mu_{xk}$ and $\sigma_x^2 = \sum_{k=1}^{2} w_k(\mu_{xk}^2 + \frac{2\sigma_{xk}^2}{N_i} - \mu_x^2)$ with $w_k = n_k/n$ [266]. Therefore, $\hat{\beta} \sim N(\mu_x, \sigma_x^2)$.

Position of Figure A1.1

Therefore, the updated mean and variance of $\hat{\beta}$, are equal to

$$\begin{aligned}
\mu_x^{'} &= w_1 \mu_{x1}^{'} + w_2 \mu_{x2} \\
&= w_1 \mu_{x1} + w_2 \mu_{x2} + w_1 \lambda \frac{\sigma_{x_1}}{\sqrt{N_1/2}} \\
&= \mu_x + w_1 \lambda \frac{\sigma_{x_1}}{\sqrt{N_1/2}} \\
\sigma_x^{'2} &= \sum_{k=1}^{2} w_k(\mu_{x_k}^2 + \frac{2\sigma_{x_k}^2}{N_k} - \mu_x^2) + D \\
&= \sigma_x^2 + D
\end{aligned} \tag{6.6}$$

where $D = w_1\Big((2\sigma_1^2/N_1)\zeta + A^2(1 - w_1^2 - w_2^2) + 2A(\mu_{x_1} - \mu_x)\Big)$, $A = (\sigma_1/\sqrt{N_1/2})\lambda$ and $\zeta = a\lambda - \big(\lambda\big)^2$.

A bias is introduced after combining the Phase II and III trial short-term outcome effect estimates as $\frac{\sigma_{x_1}\lambda \cdot w_1}{\sqrt{N_1/2}}$ [266]. Then based on equation eq1 and assuming that $\sigma_x = \sigma_y = 1$, the bias of B equals to

$$Bias(B) = \frac{w_1 \lambda \, \rho \sigma_{x_1}}{\sqrt{N_1/2}} \tag{6.7}$$

Appendix - Derivation of $Var(\hat{B})$ (A2.2)

The variance of long-term outcome B is equal to [250],

$$Var(\hat{B}) = var(\hat{b}) + \gamma^2 var(\hat{\beta}) + \beta^2 var(\hat{\gamma}) + 2\beta cov(\hat{b}, \hat{\gamma}) \qquad \text{(eqA1)}$$

An estimate of $Var(\hat{B})$ can be obtain via estimates of the relevant parameters which can be obtained directly via the regression of $X_i|t_i$ and $Y_{i2}|X_{i2}, t_{i2}$.

Assuming that t_{ik} is an indicator variable and N_k corresponds to the total sample size of the k^{th} trial, the q-dependent variance of $(\hat{\alpha}, \hat{\beta})$ can be derived as $\sigma_x^{'2}(T^{'}T)^{-1}$, where X is the design matrix of $X_i|t_i$ as follows

$$(T^{'}T)^{-1} = \begin{pmatrix} 2N_1 & N_1 \\ N_1 & N_1 \end{pmatrix}^{-1} = \begin{pmatrix} (1/N_1) & -(1/N_1) \\ -(1/N_1) & (2/N_1) \end{pmatrix} = \frac{1}{n}\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \qquad \text{(eqA2)}$$

and as a mixture of two distributions $\sigma_x^{'2} = \sigma_x^2 + D$

From eqA2, $var(\beta) = \frac{2\sigma_x^{'2}}{N}$, an estimate of which can be derived as $\hat{var}(\hat{\beta}) = \frac{2s_x^{'2}}{N}$, where s_x^2 follows from the regression of $X_i|t_i$.

Subsequently, the variance of $(\hat{a}, \hat{b}, \hat{\gamma})$ can be derived as $\sigma_o^2(1-\rho^2)(T^{'}T)^{-1}$, where T is the design matrix of $Y_{2i}|X_{2i}, t_{2i}$.

$$E(T^{'}T) = E\begin{pmatrix} 2N_2 & N_2 & \sum\limits_{C,E} x_{2i} \\ N_2 & N_2 & \sum\limits_{E} x_{2i} \\ \sum\limits_{C,E} x_{2i} & \sum\limits_{E} x_{2i} & \sum\limits_{C,T} x_{2i}^2 \end{pmatrix} = n_{III} \begin{pmatrix} 2 & 1 & \mu_C + \mu_E \\ 1 & 1 & \mu_E \\ \mu_C + \mu_E & \mu_E & 2\sigma_{2z}^2 + (\mu_C^2 + \mu_E^2) \end{pmatrix} \tag{eqA3}$$

The variance estimates are derived by inverting matrix eqA3 and replacing σ_o^2 with $\sigma_y^{'2} = \sigma_0^2 + \gamma^2 \sigma_x^{'2}$ [253].

$$var(\hat{\gamma}) = \frac{\sigma_y^{'2}(1-\rho^2)}{2N_2\sigma_x^{'2}} \tag{eqA4}$$

$$var(\hat{b}) = \frac{2\sigma_y^{'2}(1-\rho^2)}{N_2} + \frac{\sigma_y^{'2}(1-\rho^2)}{2N_2\sigma_x^{'2}}\beta^2 \tag{eqA5}$$

$$cov(\hat{b}, \hat{\gamma}) = -\frac{\sigma_y^{'2}(1-\rho^2)}{2N_2\sigma_x^{'2}}\beta \tag{eqA6}$$

Replacing eqA4, eqA5 and eqA6 in eqA1 we obtain $var(\hat{B})$

$$\begin{aligned} Var(\hat{B}) &= var(\hat{b}) + \gamma^2 var(\hat{\beta}) + \beta^2 var(\hat{\gamma}) + 2\beta cov(\hat{b}, \hat{\gamma}) \\ &\quad \frac{2\sigma_y^{'2}(1-\rho^2)}{N_2} + \frac{\sigma_y^{'2}(1-\rho^2)}{2N_2\sigma_x^{'2}}\beta^2 + \frac{2\sigma_y^{'2}\rho^2}{N} + \\ &\quad \beta^2\frac{\sigma_y^{'2}(1-\rho^2)}{2N_2\sigma_x^{'2}} + 2\beta(-\frac{\sigma_y^{'2}(1-\rho^2)}{2N_2\sigma_x^{'2}}\beta) \\ &= 2\sigma_y^{'2}\Big(\frac{(1-\rho^2)}{N_2} + \frac{\rho^2}{N}\Big) \\ \sigma_y^{'2} &= \sigma_y^2 + \gamma^2 D \end{aligned}$$

Appendix - Derivation of $MSE(\hat{B})$ (A2.3)

Based on the calculated alternative variance of the overall long-term effect $Var(\hat{B})$ and the method of moments, the $MSE(\hat{B})$ is given by

$$
\begin{aligned}
MSE(\hat{B}) &= Bias(\hat{B})^2 + Var(\hat{B}) \\
&= \Big(\frac{w_1 \lambda \, \rho \sigma_y^{'} \sigma_{x_1}}{\sigma_x^{'} \sqrt{N_1/2}}\Big)^2 + 2\sigma_y^{'2}\Big(\frac{1-\rho^2}{N_2} + \frac{\rho^2}{N}\Big) \\
&= 2\sigma_y^{'2}\Big(\frac{w_1 \rho \lambda \sigma_{x_1}}{\sigma_x^{'} \sqrt{N_1}}\Big)^2 + 2\sigma_y^{'2}\Big(\frac{1-\rho^2}{N_2} + \frac{\rho^2}{N}\Big) \\
\sigma_y^{'2} &= \sigma_y^2 + \gamma^2 D
\end{aligned}
$$

Chapter 6

Supplementary material can be found at figshare.com/s/fa5b0f8059b392f34a8a - 10.6084/m9.figshare.11977791 or/and at the online manuscript.

Chapter 7

Participants′ outcomes gone missing within a network of interventions: Bayesian modeling strategies

L Spineli
C Kalyvas
K Pateras

Statistics in Medicine. 2019 ; 38: 3861–3879.
https://doi.org/10.1002/sim.8207

Chapter 7

Abstract

To investigate the implications of addressing informative missing binary outcome data (MOD) on network meta-analysis (NMA) estimates while applying the missing at random (MAR) assumption under different prior structures of the missingness parameter. In three motivating examples, we compared six different prior structures of the informative missingness odds ratio (IMOR) parameter in logarithmic scale under pattern-mixture and selection models. Then, we simulated 1000 triangle networks of two-arm trials assuming informative MOD related to interventions. We extended the Bayesian random-effects NMA model for binary outcomes and node-splitting approach to incorporate these 12 models in total. With interval plots, we illustrated the posterior distribution of log OR, common between-trial variance (τ^2), inconsistency factor and probability of being best per intervention under each model. All models gave similar point estimates for all NMA estimates regard-less of simulation scenario. For moderate and large MOD, intervention-specific prior structure of log IMOR led to larger posterior standard deviation of log ORs compared to trial-specific and common-within-network prior structures. Hierarchical prior structure led to slightly more precise τ^2 compared to identical prior structure, particularly for moderate inconsistency and large MOD. Pattern-mixture and selection models agreed for all NMA estimates. Analyzing informative MOD assuming MAR with different prior structures of log IMOR affected mainly the precision of NMA estimates. Reviewers should decide in advance on the prior structure of log IMOR that best aligns with the condition and interventions investigated.

7.1 Introduction

Plenty of empirical studies on reporting quality of systematic reviews with conventional meta-analyses have revealed several shortcomings in the reporting and administration of missing binary outcome data (MOD) [267, 268, 269, 270]. Recommendations aiming to improve reporting of systematic reviews with regards to MOD already exist and are built upon this comprehensive empirical evidence. Contrariwise, proposed guidelines for the administration of MOD in systematic reviews have evolved in the absence of simulation studies using only intuitive argumentations [117, 271]; for example, in the Cochrane Handbook, it is stated that "*[imputing the missing data with replacement values] fails to acknowledge uncertainty in the imputed values and results, typically, in confidence intervals that are too narrow*" (see chapter 16.1.2 in the work of Higgins and Green [117]). Current directions to deal with MOD in systematic reviews include (i) analysis of observed outcomes as a primary analysis, (ii) imputation of MOD under plausible scenarios as a sensitivity analysis, and (iii) statistical modelling of missingness mechanisms (ie, reasons that triggered MOD) [117]. The first two options are the most commonly adopted in systematic reviews [267, 268, 269]. Nonetheless, they have been criticized for being employed inefficiently through data elimination or augmentation before analysis, respectively, and hence for ignoring the uncertainty induced by the scenarios considered [117, 272, 273]. In turn, these options may compromise the conclusions of the systematic review [274].

Statistical modeling of MOD has received little attention in systematic reviews with two (for example, the works of Ejere et al [275], Mayo-Wilson et al[276], and Virgili et al [277]) or more interventions (for example, the works of Watt et al [278] and Veroniki et al [279]). As opposed to imputation or exclusion, modeling MOD comprises an elegant framework that adjusts for bias due to MOD and fully acknowledges the uncertainty about the scenarios considered for the missingness mechanism. This is achieved by modelling the joint distribution of the outcomes (observed and missing) and missingness indicator [280]. This joint distribution is further factorized in two ways: a distribution of the outcome, given the missingness indicator, and a distribution of that indicator (pattern-mixture model) [281] or a distribution of the missingness indicator, given the underlying outcome, and a distribution of the underlying outcome (selection model) [282]. Selection model is more prevalent in the literature for clinical trials [283], while pattern-mixture model has been most frequently described in the

analysis of series of trials [269]. Modelling MOD using either pattern-mixture or selection models offers a thorough investigation of the underlying missingness mechanisms across different trials and interventions [273, 284, 285]. These mechanisms can be naturally explored using Bayesian approaches, where the reviewer assigns an informative prior distribution on the missingness parameter (ie, an absolute or relative measure of the relationship between outcome and missingness indicator) to indicate a specific scenario alongside the uncertainty for that scenario [273].

The existing directions on reporting and handling MOD in conventional systematic reviews are of great relevance and importance also for systematic reviews with network meta-analysis (NMA). NMA offers an in-depth exploration of the missingness mechanisms in the network as interventions may carry a different degree of and reasons for MOD in different comparisons and this information cannot be located in isolated conventional meta-analyses. Moreover, due to the addition of interventions, assumptions, and model parameters that structure this framework, addressing MOD in NMA can reveal their implications on model parameters beyond the standard meta-analytic ones. Since the statistical methodology of NMA has been refined and implemented mainly within the Bayesian framework, [116, 286, 287] we view statistical modeling with the assignment of carefully selected prior distribution on the missingness parameter as a natural way to handle MOD in a network of interventions.

To our knowledge, there is currently no published empirical or simulation study on the comparative performance of models for MOD using Bayesian approaches in terms of meta-analysis or NMA estimates. Consequently, the analyst misses the knowledge of the overall performance of models for aggregated MOD to critically decide on the proper models to apply. To shed light on this knowledge gap, we set up a comprehensive simulation study using empirical evidence from published NMAs in a wide range of health-related fields to inform the simulation setting for a triangle network of two-arm trials. This simulation study aims to designate the factors that may affect the performance of modelling informative MOD (ie, the missingness mechanism depends on the unobserved outcomes [288]) on the basis of core NMA estimates while assuming missing at random (MAR) for analysis as a starting point [272, 284, 289]. Furthermore, the simulation results supplement the observations from a relevant empirical study [290] in order to provide empirically-based recommendations for a

proper modelling of MOD in systematic reviews.

This article is organized as follows. In Section 6.2, we present the Bayesian random-effects NMA model for binary outcomes in the absence of MOD (as described by Dias et al [291]), and then, we expand the model to incorporate MOD through pattern-mixture and selection models [273, 285]. Then, we present the prior structures for the missingness parameter that we considered in the simulation study. In Section 6.3, we illustrate these prior structures under pattern-mixture and selection models in three published systematic reviews with NMA. In Section 6.4, we describe a novel simulation set-up that combines already established data generation models for conventional meta-analysis with specific algorithms to incorporate MOD in NMA, and we present the results in Section 6.5; in Section 6.6, we discuss the findings and limitations of the study and we provide recommendations, and we conclude in Section 6.7.

7.2 Missing outcome data in network meta-analysis

Bayesian random-effects NMA model

Consider a network of N trials that investigate different sets of T interventions for a specific condition. The outcome of interest is binary and the frequency of outcome in arm $k = 1, 2, \ldots, \alpha_i$ of trial $i = 1, 2, \ldots, N$ is assumed to be a realization from the binomial distribution

$$r_{i,k} \sim Bin(p_{i,k}, n_{i,k})$$

with $p_{i,k}$ being the underlying risk of an event (the parameter of interest) and $n_{i,k}$ the randomized sample in arm k of trial i. Then, using a logit function, as described by Dias et al [291], the log odds of event in arm k of trial i are defined as follows:

$$logit(p_{i,k}) = u_i + \theta_{i,k1} \tag{7.1}$$

$u_i = logit(p_{i,1})$ is the log odds of event in the baseline arm of trial i and $\theta_{i,k1}$ is the log odds ratio (OR) of event in arm k relative to the baseline arm that typically follows a normal distribution with mean $\mu_{t_{i,k}t_{i,1}}$ and variance τ^2 commonly assumed to be constant across different comparisons. Index $t_{i,k}$ indicates the intervention studied in arm k of trial i.

Incorporating multi-arm trials

In a trial i with $a_i > 2$ arms, log ORs are correlated since they share the same comparator, and therefore, the vector θ_i of $\alpha_i - 1$ log ORs follows a multivariate normal distribution [291, 292]

$$\theta_i = \begin{pmatrix} \theta_{i,21} \\ \vdots \\ \theta_{i,\alpha_i 1} \end{pmatrix} \sim MVN_{\alpha_i - 1} \left[\begin{pmatrix} \mu_{t_{i,2}t_{i,1}} \\ \vdots \\ \mu_{t_{i,\alpha_i}t_{i,1}} \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2/2 & \vdots & \tau^2/2 \\ \tau^2/2 & \tau^2 & \vdots & \tau^2/2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2/2 & \tau^2/2 & \vdots & \tau^2 \end{pmatrix} \right]$$

which, under the consistency assumption, is equivalent to conditional univariate normal distributions as follows [291]:

$$\theta_{i,k1} | \begin{pmatrix} \theta_{i,21} \\ \vdots \\ \theta_{i,(\alpha_i - 1)1} \end{pmatrix} \sim N \left[(\mu_{t_{i,k}A} - \mu_{t_{i,1}A}) + \frac{1}{\alpha_i} \sum_{j=2}^{\alpha_i - 1} (\theta_{i,j1} - \mu_{t_{i,j}A} - \mu_{t_{i,1}A}), \frac{\alpha_i}{2 \cdot (\alpha_i - 1)} \cdot \tau^2 \right]$$

where μ_{tA} reflects the relative treatment effects of the comparisons with the reference intervention of the network, A (known as basic parameters [293]). Then, using the consistency equation, the relative treatments effects of all possible non-reference comparisons can be obtained as functions of the basic parameters

$$\mu_{tl} = \mu_{tA} - \mu_{lA},$$

with $t, l = \{B, C, \ldots, T\} \ngeq A$ and $t \neq l$.

In the Bayesian framework, all parameters of the model are random variables that need proper prior distributions. In the present study, we used non-informative normal prior distribution with mean 0 and variance 10000 for the location parameters (ie, u_i and μ_{tA}), whereas we considered $HN(0, 1)$ (median: 0.98, interquartile range [IQR]: 0.51-1.96) as a weakly informative prior distribution on τ due to trial sparsity in the investigated networks that may compromise a proper estimation of τ.

Rank probabilities for each intervention

To facilitate decision-making, we can estimate for each intervention the probability of being first, second, third, and so on for a specific outcome [294]. These rank probabilities are estimated by ordering the basic parameters in each iteration of the Markov chain Monte Carlo (MCMC) simulation and then, for each intervention, calculating the frequency to achieve a specific rank out of the number of iterations.

Node-splitting approach to assessing local inconsistency

To assess possible inconsistency locally while using the whole network to obtain an indirect effect for a comparison of a closed loop, Dias et al [295] proposed the node-splitting approach within a Bayesian framework. Specifically, a comparison from a closed loop is isolated (split) and random-effects meta-analysis is applied, whereas the remaining network is used to estimate an indirect effect for the split comparison. Then, the difference between direct and indirect effect for that comparison yields a posterior distribution for the inconsistency between these two effects, known as inconsistency factor (IF). A large posterior probability of IF being different from zero (eg, above 95%) provides sufficient evidence that inconsistency may be present in a specific loop. To improve the estimation of τ^2, a common τ^2 is assumed for both meta-analysis and NMA model after removing the trials of the split comparison.

Modelling missing outcome data

Pattern-mixture model

Suppose that $m_{i,k}$ participants were missing (for reasons related or not to the design and conduct of the trial) in arm k of trial i with probability $q_{i,k}$, whereas among those $n^o_{i,k} = n_{i,k} - m_{i,k}$ participants who were observed, only $r^o_{i,k}$ experienced the studied outcome with probability $p^o_{i,k}$. It follows that the number of MOD and the number of observed events in arm k of trial i are realizations from the respective binomial distributions

$$m_{i,k} \sim Bin(q_{i,k}, n_{i,k}) \text{ and } r^o_{i,k} \sim Bin(p^o_{i,k}, n^o_{i,k})$$

In the presence of MOD, a pattern-mixture model can be considered, where $p_{i,k}$ is modelled conditional on whether the underlying event is observed or missing

$$p_{i,k} = p^o_{i,k} \cdot (1 - q_{i,k}) + p^m_{i,k} \cdot q_{i,k} \tag{7.2}$$

where $p^m_{i,k}$ is the missingness parameter and indicates the probability of event conditional on MOD in arm k of trial i. The parameters $p^o_{i,k}$ and $q_{i,k}$ can be estimated directly from the data, whereas we need a proper prior distribution on $p^m_{i,k}$ to describe a plausible missingness mechanism.

Following thework of Turner et al [273] after rearranging Equation (7.2) to link $p^o_{i,k}$ with the remaining parameters, we obtain the following:

$$p^o_{i,k} = \frac{p_{i,k} - p^m_{i,k} \cdot q_{i,k}}{1 - q_{i,k}}$$

Subsequently, we use Equation (7.1) with a random-effects model for $\theta_{i,k1}$ to apply the NMA model.

Selection model

Instead of applying separate binomial distributions, we can jointly model all observed data via the following multinomial distribution [284, 285]:

$$L_{1,i,k} \sim M(p_{1,i,k}, p_{2,i,k}, p_{3,i,k}, n_{i,k})$$

where $L_{1,i,k}$ is a vector of all data observed in arm k of trial i, namely, $(r^o_{i,k}, n_{i,k} - r^o_{i,k} - m_{i,k}, m_{i,k})^T$ and

$$\begin{aligned} p_{1,i,k} =& (1 - c_{1,i,k}) \cdot p_{i,k} \\ p_{2,i,k} =& (1 - c_{0,i,k}) \cdot (1 - p_{i,k}) \end{aligned}$$

$$q_{1,i,k} = p_{3,i,k} = c_{1,i,k} \cdot p_{i,k} + c_{0,i,k} \cdot (1 - p_{i,k}) \tag{7.3}$$

where $p_{1,i,k}$ reflects the marginal probability of observing the underlying event, $p_{2,i,k}$ reflects the marginal probability of observing the underlying non-event, and $q_{i,k}$ is the probability of MOD out of the randomized sample in arm k of trial i, respectively. The latter equation actually describes the selection model that has already been proposed in a conventional meta-analysis [284] and extended to operate in NMA [285]. Then, parameters $c_{1,i,k}$ and $c_{0,i,k}$ indicate the probability of MOD conditional on those participants with the underlying event and the probability of MOD conditional on those participants without the underlying event, respectively. Apart from $q_{i,k}$, all other parameters are not estimable from the data, and hence, we need to assign proper prior distributions for precise inference to be possible.

Informative missingness odds ratio as missingness parameter

In the present study, we focus on the informative missingness odds ratio (IMOR) parameter, which, under the pattern-mixture model, is defined as follows [272, 273, 296]:

$$\delta_{i,k}^{PM} = \frac{p_{i,k}^{m}/(1-p_{i,k}^{m})}{p_{i,k}^{o}/(1-p_{i,k}^{o})}$$

while under the selection model, it is defined as [284, 285]

$$\delta_{i,k}^{S} = \frac{c_{1,i,k}/(1-c_{1,i,k})}{c_{0,i,k}/(1-c_{0,i,k})}$$

Similar to OR, IMOR takes non-negative values; nevertheless, due to different factorizations of the same joint distribution of outcome and missingness indicator under pattern-mixture (PM) and selection (S) models, IMOR has different interpretation with respect to these models:

- $\delta_{i,k}^{PM} > 1$, the odds of underlying event among those participants being missing is more likely than the odds of underlying event among those participants being observed in arm k of trial i;
- $\delta_{i,k}^{S} > 1$, the odds of MOD among participants with underlying event is more likely than the odds of MOD among participants without underlying event in arm k of trial i;
- $\delta_{i,k}^{PM} < 1$, the odds of underlying event among those participants being observed is more likely than the odds of underlying event among those participants being missing in arm k of trial i;
- $\delta_{i,k}^{S} < 1$, the odds of MOD among participants without underlying event is more likely than the odds of MOD among participants with underlying event in arm k of trial i;
- $\delta_{i,k}^{PM} = 1$, the outcome is similarly distributed between those participants being missing and those being observed in arm k of trial i (ie, MAR assumption);
- $\delta_{i,k}^{S} = 1$, MOD are equally likely to occur among participants with underlying event and those without underlying event in arm k of trial i (ie, MAR assumption).

Like OR, IMOR is applied in the logarithmic scale but it is back-transformed to facilitate in the interpretation

$$log(\delta^{PM}_{i,k}) = \phi^{PM}_{i,k} = logit(p^{m}_{i,k}) - logit(p^{o}_{i,k})$$
$$log(\delta^{S}_{i,k}) = \phi^{S}_{i,k} = logit(c_{1,i,k}) - logit(c_{0,i,k})$$

under pattern-mixture model and selection model, respectively.

Structural assumptions to model informative missingness odds ratio

To investigate the underlying missingness mechanisms while acknowledging the uncertainty regarding our prior belief, normal prior distributions are assigned on $\phi^{l}_{i,k}$ with carefully selected values for the mean ($\mu^{\phi}_{i,k}$) and variance ($\sigma^{2}_{i,k}$) that reflect a plausible belief about the missingness mechanism on average and make $\phi^{l}_{i,k}$ identifiable, respectively,

$$\phi^{l}_{i,k} \sim N(\mu^{\phi}_{i,k}, \sigma^{2}_{i,k}) \text{ for } l = PM, S$$

Following the work of White et al [284], we considered $\phi^{l}_{i,k}$'s to be on average MAR (as recommended by relevant published literature to address MOD in the primary analysis [272, 284, 288]) and exchangeable across trials and interventions, that is, $\mu^{\phi}_{i,k} = 0$ and $\sigma^{2}_{i,k} = \sigma^{2}$. White et al [284, 296], recommended choosing $\sigma^{2} \in [0.25, 4]$, which covers a range of values for log IMOR reflecting liberal to conservative uncertainty about the missingness scenario considered. In the present study, we used $\sigma^{2} = 1$:

$$\phi^{l}_{i,k} \sim N(0, 1) \text{ for } l = PM, S \tag{7.4}$$

The prior distribution (4) can be shaped further to accommodate our prior beliefs regarding how different $\phi^{l}_{i,k}$'s can be related within the network [273, 284]. Following our empirical study [290], we considered identical and hierarchical prior structure for $\phi^{l}_{i,k}$. Under identical structure, $\phi^{l}_{i,k}$ is assumed to be the same across trials that investigate the same interventions

but different across interventions (intervention-specific)

$$\phi^l_{i,k} = \phi^l_{t_{i,k}}, \phi^l_{t_{i,k}} \sim N(0,1)$$

or the same across interventions compared in a trial but different across trials (trial-specific)

$$\phi^l_{i,k} = \phi^l_i, \phi^l_i \sim N(0,1)$$

or identical across all trials and interventions (common-within-network)

$$\phi^l_{i,k} = \phi^l, \phi^l \sim N(0,1)$$

Hierarchical structure "relaxes" the identical structure by assuming $\phi^l_{i,k}$'s to be different yet related to each other. Then, intervention-specific $\phi^l_{i,k}$ under on average MAR is defined as

$$\phi^l_{i,k} \sim N(\mu^\phi_{t_{ik}}, \sigma^2_{t_{ik}}) \text{ with } \mu^\phi_{t_{ik}} \sim N(0,1), \sigma^2_{t_{ik}} \sim U(0,1)$$

trial-specific $\phi^l_{i,k}$ on average MAR is defined as

$$\phi^l_{i,k} \sim N(\mu^\phi_i, \sigma^2_i) \text{ with } \mu^\phi_i \sim N(0,1), \sigma^2_i \sim U(0,1)$$

and common-within-network $\phi^l_{i,k}$ on average MAR is defined as

$$\phi^{l}_{i,k} \sim N(\mu^{\phi}, \sigma^2) \text{ with } \mu^{\phi} \sim N(0,1), \sigma^2 \sim U(0,1)$$

We assigned a uniform distribution on σ, σ_i, and $\sigma_{t_{ik}}$; however, other appropriate prior distributions for variance components can be also considered [144, 160].

7.3 Illustrative examples

Example 1: low missing outcome data

Bottomley et al [297] investigated the effectiveness of seven interventions measured as the investigator's global assessment response at 4 weeks in patients with moderately severe scalp psoriasis. A total of 9 trials (7 two-arm, 1 three-arm, and 1 four-arm trials) with 5889 patients (median per trial: 237, IQR: 136-419) formed the network (Figure 7.1A). For this outcome, MOD were low (median per trial: 3%, IQR: 1%-6%) in the included trials. Positive log OR indicated a beneficial effect of the first intervention of the comparison.

Overall, results on log ORs were almost identical for all missingness models (pattern-mixture or selection model) and prior structures of log IMOR (Supporting Information S.2; Figure S1). As a result, the ranking curves were indistinguishable for different prior structures of log IMOR in both missingness models (Supporting Information S.2; Figure S2). Results were also similar for τ^2, although the 95% credible intervals (CrIs) were slightly narrower for hierarchical, trial-specific prior structure of log IMORs in both missingness models (Supporting Information S.2; Figure S1). Results on node-splitting were in line with those on basic parameters (Supporting Information S.2; Figure S3).

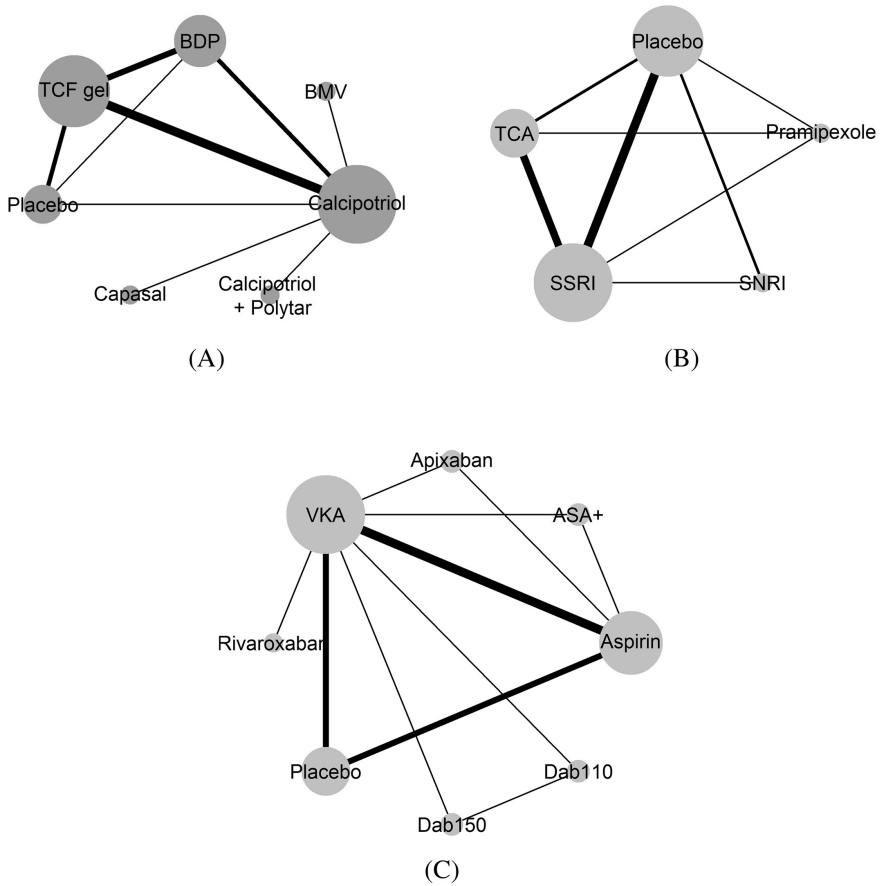

Figure 7.1: A series of network plots on (A) the effectiveness of topical therapies for moderately severe scalp psoriasis [297], (B) the efficacy of antidepressants in Parkinson's disease [298], and (C) the prevention of a stroke episode in patients with atrial fibrillation using oral antithrombotics [299]. The thickness of the lines and the size of the nodes are proportional to the number of trials and the number of patients randomized in the respective treatments, respectively. ASA+, aspirin plus clopidogrel; Dab110, dabigatran 110 mg; Dab150, dabigatran 150 mg; BDP, betamethasone dipropionate; BMV, betamethasone valerate; SNRI, serotonin–norepinephrine reuptake inhibitor; SSRI, selective serotonin reuptake inhibitor; TCA, tricyclic antidepressant; TCF, two-compound formulation; VKA, vitamin K antagonist.

Example 2: moderate and balanced missing outcome data

Liu et al [298] assessed the comparative effectiveness of four antidepressants and placebo in Parkinson's disease measured as the proportion of patients who had a reduction of at least

50% from the baseline score (Figure 7.1B). For this outcome,the authors included a total of 11 trials (8 two-arm and 3 three-arm trials) with 801 patients (median per trial: 19, IQR: 17-33). MOD were moderate (median per trial: 16%, IQR: 12%-24%) and balanced (median per trial: 4%, IQR: 2%-11%) in the included trials. Positive log OR indicated beneficial effect of the first intervention of the comparison.

Results on log ORs were similar overall, albeit the 95% CrIs were slightly wider for (identical and hierarchical) intervention-specific prior structure of log IMORs in both missingness models (Supporting Information S.3; Figure S4). Nevertheless, τ^2 was slightly lower (and with slightly narrower 95% CrIs) for hierarchical as compared to identical prior structure of log IMOR regardless of further structural assumptions or missingness model. No profound differences were observed on rank probabilities (Supporting Information S.3; Figure S5) and the results from node-splitting approach (Supporting Information S.3; Figure S6).

Example 3: moderate and unbalanced missing outcome data

Dogliotti et al [299] assessed the comparative effectiveness of seven antithrombotic therapies and placebo in terms of preventing a stroke episode in patients with atrial fibrillation (Figure 7.1C). The authors included 16 trials (12 two-arm and 4 three-arm trials) with 79808 patients (median per trial: 391, IQR: 211-2940). MOD were moderate (median per trial: 19%, IQR: 13%-23%) and slightly unbalanced (median per trial: 7%, IQR: 3%-10%). Negative log OR indicated a beneficial effect of the first intervention in the comparison. Different assumptions about the prior structure of log IMOR appeared to implicate mostly on the width of 95% CrIs for all NMA estimates.

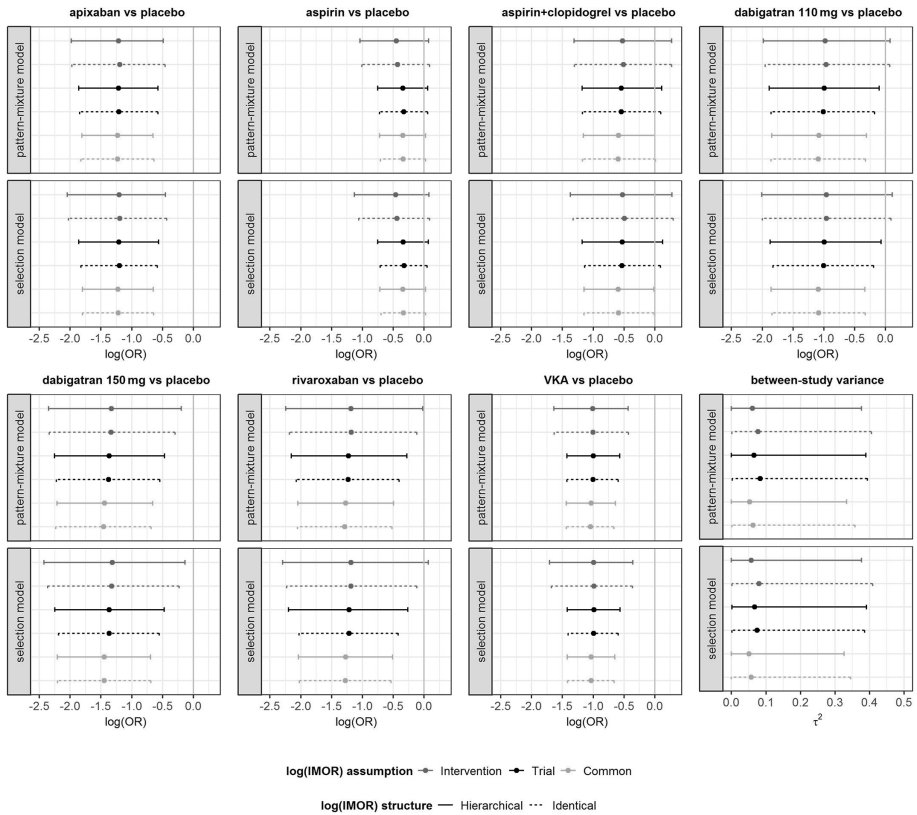

Figure 7.2: Interval plots on log ORs for basic parameters (posterior mean and 95% credible interval) and between-trial variance (τ^2; posterior median and 95% credible interval) when there are moderate and unbalanced missing outcome data (MOD) in the network [299]. Results are compared in terms of model for MOD (pattern-mixture, model selection model), structure (hierarchical, identical), and assumption (intervention-specific, trial-specific, common-within-network) for prior normal distribution on log IMOR assuming missing at random. IMOR, informative missingness odds ratio; OR, odds ratio.

Overall, intervention-specific prior of log IMOR led to wider 95% CrIs for log ORs in both missingness models, whereas common-within-network prior led to narrower 95% CrIs for log ORs to some extent. In fact, 95% CrI for log ORs were slightly wider under hierarchical than identical structure. Consequently, the superiority of dabigatran at 110 mg and rivaroxaban against placebo turned into inconclusive when log IMOR was assumed to have intervention-specific prior structure (Figure 7.2). Furthermore, τ^2 was relatively lower and slightly more precise under hierarchical structure, especially, for common-within-network log IMORs.

Since, the common-within-network structure provided the narrowest 95% CrIs for logORs, it led to relatively larger rank probabilities as opposed to intervention-specific prior structure, especially for aspirin, aspirin plus clopidogrel, and VKA (Figure 7.3). Results on node-splitting were in line with those on basic parameters (Supporting Information S.4; Figure S7).

7.4 Simulation setting

Data generation without missing outcome data

We simulated a triangle network of two-arm trials and three interventions: placebo, new intervention, and old intervention. The comparison of interest was new versus old intervention. We assumed a typical loop like that in the work of Veroniki et al [200] with four trials for old intervention versus placebo, three trials for new intervention versus placebo, and one trial for new versus old intervention. To determine the sample size in each arm of every trial, we used information directly from the networks that we collected in our previous empirical work [288]. For each trial, we considered equally sized arms with sample size generated from a uniform distribution with support in the range defined by the second and third quartile of the arm sizes (Supporting Information S.5; Figure S8(a))

$$n^{E}_{i,k} = n^{C}_{i,P} \sim U(102, 187), k = N, O \text{ (placebo-controlled trials)}$$

$$n^{E}_{i,N} = n^{C}_{i,O} \sim U(128, 241) \text{ (old-controlled trials)}$$

where N, O, and P stand for new intervention, old intervention and placebo, respectively, whereas E and C stand for experimental and control arm, respectively. We considered a binary (beneficial) outcome measured in the logOR scale. We assumed $\mu_{NP} = log(2)$ and $\mu_{OP} = log(1.5)$ to be the underlying log OR for new and old intervention against placebo, respectively, whereas we obtained the underlying log OR for new versus old intervention through the consistency equation

$$\mu_{NO} = \mu_{NP} - \mu_{OP} + IF$$

with IF being sampled from the t-distributions $t(\mu = 0, \sigma^2 = 0.442, df = 3)$ and $t(\mu = 1, \sigma^2 = 0.442, df = 3)$ to reflect low and moderate inconsistency on average, respectively, according to our empirical work (Supporting Information S.5; Figure S8(b)) [288].

We generated the number of events in each arm of every trial using the data-generating model (DGM) described by Hartung and Knapp for a random-effects pairwise meta-analysis [198, 241]. The description of this DGM is available as in Supporting Information (S.6). Using information from our network collection [288], initial event risks for the control arms were generated from a uniform distribution with support in the range defined by the second and third quartile of the event risks (Supporting Information S.5; Figure S8(c))

$$p_{i,P}^{C,0} \sim U(0.27, 0.40) \text{ and } p_{i,O}^{C,0} \sim U(0.63, 0.76)$$

for placebo-controlled and old-controlled trials, respectively.

We incorporated τ^2 (assumed common-within-network) in the DGM assuming smaller variability in log odds for placebo (Supporting Information S.5; Figure S8(d)) but equal in log odds for active arms, respectively. In terms of scenarios for τ^2, we selected the predictive log-normal distributions $LN(-3.95, 1.342)$ (median: 0.02; IQR: 0.01-0.04) and $LN(-2.56, 1.742)$ (median: 0.08; IQR: 0.03-0.26) to reflect small and substantial τ^2, respectively. These predictive distributions referred to the expected τ^2 in a future meta-analysis for all-cause mortality and a generic healthcare setting, respectively [153]. Finally, we generated the true probability of being best for each intervention by ordering the simulated true log ORs of placebo comparisons as generated from the normal distribution $N(\mu_{kP}, \tau^2)$ with $k = N, O$ and then calculating the number of times each intervention ranked first out of the total simulations.

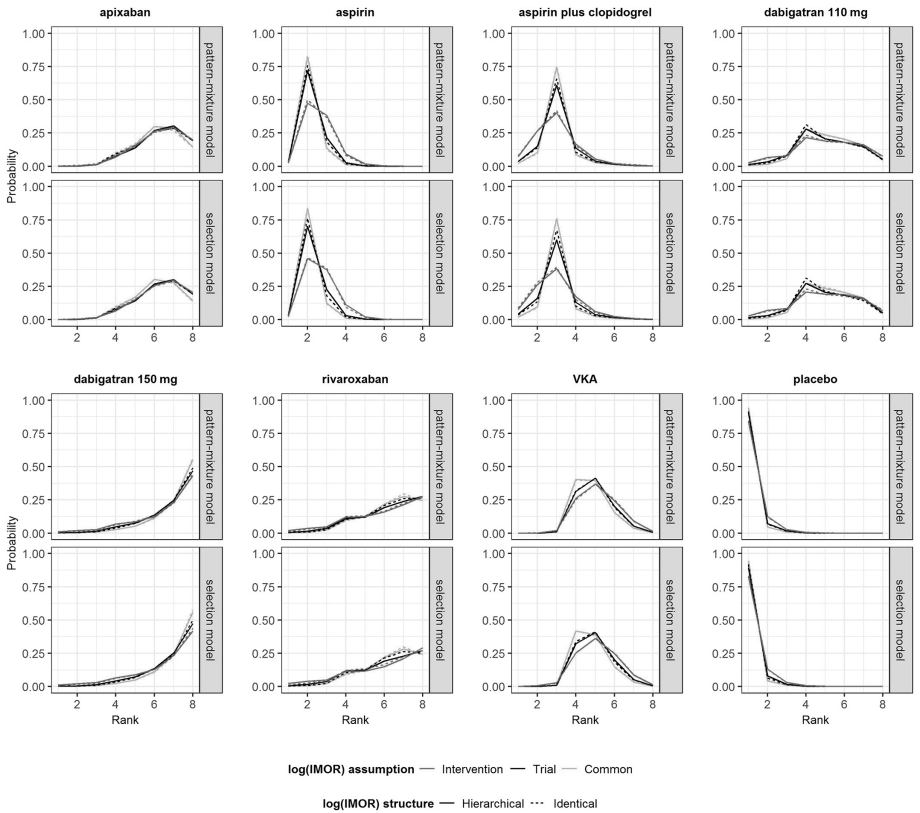

Figure 7.3: Rankograms of seven interventions when there are moderate and unbalanced missing outcome data (MOD) in the network [299]. Posterior mean rank probabilities are compared in terms of model for MOD (pattern-mixture model, selection model), structure (hierarchical, identical) and assumption (intervention-specific, trial-specific, common-within-network) for prior normal distribution on log IMOR under missing at random. IMOR, informative missingness odds ratio.

Data generation while incorporating missing outcome data

Following the motivating examples (Section 7.3), we focused only on moderate and large MOD as they affected the performance of the modelling strategies to some extent, contrary to low MOD. Note that, under low MOD, we found that all modelling strategies had almost the same performance for log OR, IF and probability of being best but similar performance for τ^2 (results not shown). To ensure balance in MOD between the compared arms, we generated %MOD in the experimental arm, $q^E_{i,k}$ with $k = N, O$, from $U(0.05, 0.20)$ and $U(0.21, 0.40)$ to

indicate moderate and large MOD, respectively (in line with the "five-and-twenty rule," as described in Supporting Information S.1), and we considered $q^C_{i,P} = q^E_{i,k}$ with $k = N, O$ and $q^C_{i,O} = q^E_{i,N}$ for the control arms in placebo-controlled and old-controlled trials, respectively. In another scenario, to capture the imbalance in MOD between the compared arms, we assumed placebo to have more MOD than the active arms following our empirical study (Supporting Information S.5; Figure S8(e)) and old intervention to have more MOD in the old-controlled trials [290]. Details on the generation of unbalanced MOD are available as in Supporting Information (S.7).

Then, we generated the number of MOD in each arm of every trial through the following binomial distributions:

$$m^E_{i,k} \sim Bin(q^E_{i,k}, n^E_{i,k}), k = N, O$$
$$m^C_{i,k} \sim Bin(q^C_{i,k}, n^C_{i,k}), k = O, P$$

for the experimental and control arm, respectively. We used intervention-specific log IMORs under the pattern-mixture model to indicate the outcome among the missing participants in each arm of every trial. Specifically, for each trial, we assumed patients randomized in the new or old intervention to be twice more likely to be missing due to the improvement of their outcome as opposed to patients receiving placebo. We considered $\sigma^2 = 1$ for the variance of log IMORs. As another scenario, we assumed MAR on average (ie, $\mu^{\phi}_{i,k} = 0$) with $\sigma^2 = 1$. Details on the generation of log IMORs are available as in Supporting Information (S.8). Then, we used the linkage function as described by Turner et al [273] (equation 7, there) to obtain the probability of events given observed outcomes, $p^{E,obs}_{i,k}$ and $p^{C,obs}_{i,k}$ in arm k of trial i for the experimental and control arm, respectively. The formula to obtain the probability of observed events in each arm is available as in Supporting Information (S.9).

Table 7.1: Scenarios for the simulation setup. Note: C: control; E: experimental arm; IF: inconsistency factor; IMOR: informative missingness odds ratio; LOR: log odds ratio; N: new intervention; O: old intervention; P: placebo. Typical loop as defined by Veroniki et al [200]. Using predictive log-normal distributions that correspond to all-cause mortality and generic health setting for small and substantial between-trial variance, respectively [153].

Number of trials per comparison	
$NO = 1$, $NP = 3$, $OP = 4$	
Trial size $(n^E_{i,k} = n^C_{i,k}) = n_i$ **in trial** i	
Placebo-controlled trials	$n_i \sim U(102, 187)$
Old-controlled trials	$n_i \sim U(128, 241)$
Initial event rates of control arm in trial i	
Placebo-controlled trials	$p^{C,0}_{i,P} \sim U(0.27, 0.40)$
Old-controlled trials	$p^{C,0}_{i,O} \sim U(0.63, 0.76)$
Balanced risk of missing outcome data $(q^E_{i,k} = q^C_{i,k} = q_i$ **in trial** $i)$	
Moderate	$q_i \sim U(0.05, 0.20)$
Large	$q_i \sim U(0.21, 0.40)$
Unbalanced risk of missing outcome data $(q^E_{i,k} < q^C_{i,k}$ **in trial** $i)$	
Moderate	$q^E_i \sim U(0.05, 0.10), q^C_i \sim U(0.11, 0.20)$
Large	$q^E_i \sim U(0.21, 0.30), q^C_i \sim U(0.31, 0.40)$
Missing mechanisms via log (IMOR)	
Informative	$\phi_{i,P} \sim TN(\mu = -log(2), \sigma^2 = 1, \alpha = log(1))$
	$\phi_{i,k} \sim TN(\mu = log(2), \sigma^2 = 1, \alpha = log(1)), k = N, O$
Missing at random	$\phi_{i,k} \sim N(0, 1), k = N, O, P$
Treatment effects	
Basic parameters	$LOR_{NP} = log(2), LOR_{OP} = log(1.5)$
Functional parameters	$LOR_{NO} = LOR_{NP} - LOR_{OP} + IF$
Loop inconsistency	
Inconsistency factor (IF)	$IF \sim t(\mu = 0, \sigma^2 = 0.44^2, df = 3)(low)$
	$IF \sim t(\mu = 1, \sigma^2 = 0.44^2, df = 3)(moderate)$
Common between-trial variance	
Predictive distribution	$\tau^2 \sim LN(-3.95, 1.34^2)(small)$
	$\tau^2 \sim LN(-2.56, 1.74^2)(moderate)$
Probability of being best	
New intervention	93% and 76% for small and substantial τ^2, respectively
Old intervention	7.3% and 24% for small and substantial τ^2, respectively
Placebo	0% and 0.1% for small and substantial τ^2, respectively

Finally, we generated the number of events given the observed outcomes in each arm of every trial as follows:

$$r_{i,k}^{E,obs} \sim Bin(p_{i,k}^{E,obs}, n_{i,k}^{E} - m_{i,k}^{E}), k = N, O$$
$$r_{i,k}^{C,obs} \sim Bin(p_{i,k}^{C,obs}, n_{i,k}^{C} - m_{i,k}^{C}), k = O, P$$

for the experimental and control arm, respectively. Table 1 summarizes all simulation scenarios considered in the present study.

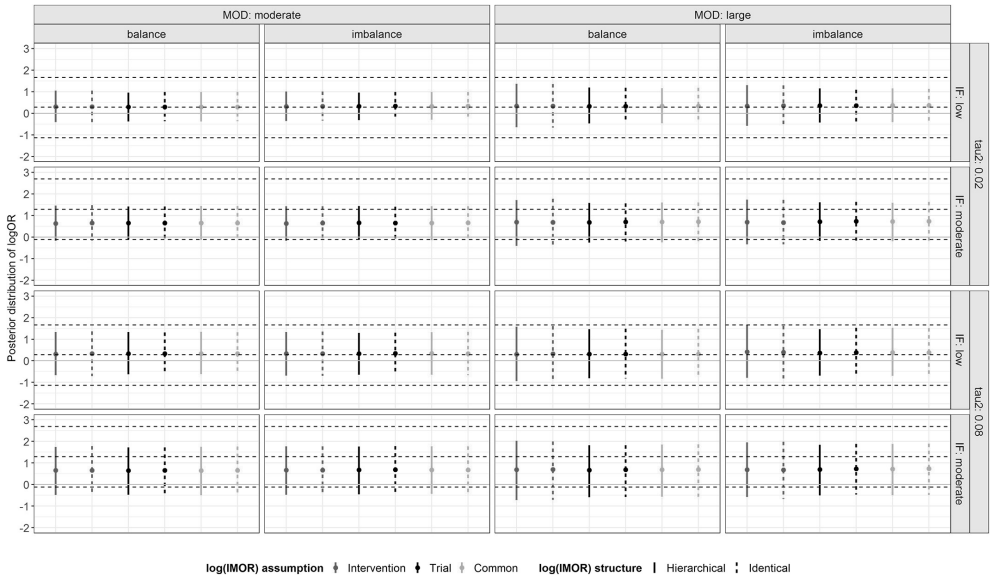

Figure 7.4: Posterior distribution of log OR (between new and old intervention) under informative missingness while using pattern-mixture model and accounting for the extent of missing outcome data (moderate, large), balance of missing outcome data (balance, imbalance), extent of τ^2 (small, substantial), and extent of inconsistency (low, moderate). The horizontal dotted lines reflect the 95% interval and mean of the simulated distribution of log OR under low and moderate true inconsistency. IF, inconsistency factor; MOD, missing outcome data.

Results presentation and model specification

For each scenario, we simulated 1000 triangle networks and, for each scenario, we evaluated the posterior distribution of μ_{NO}, τ^2, IF and probability of being best for each intervention. For each NMA estimate, we used interval plots to present the simulation results in order to fully reflect the dispersion of the results for each scenario. We decided to present in the main text only results on prior structures of log IMOR under pattern-mixture model as it is the most frequently reported model in systematic reviews [269]. Results on prior structures of log IMOR under selection model are available in Supporting Information (S.11; Figures S10-S13). Furthermore, we focused on informative MOD with moderate and large extent for being the most plausible scenarios in a medical setting. Results on prior structures of log IMOR when MOD are MAR can be found in Supporting Information (S.12; Figures S14-S17). Simulations and analyses were performed in the line with the motivating examples (Section 7.3). For each of the 1000 simulations, thinning equal to 3 was used for 20 000 updates and a burn-in of 2000 MCMC samples [300].

7.5 Results

Posterior distribution of log OR (μ_{NO})

Under low inconsistency, the posterior mean of log OR almost converged with the simulated distribution for all prior structures of log IMOR regardless of extent and balance of MOD (Figure 7.4). Credible intervals were broadly similar for moderate MOD. Subtle differences in the CrIs were observed for large MOD: assuming intervention-specific log IMORs led to slightly wider CrIs (similarly for identical and hierarchical structure) compared to trial-specific and common-within-network prior structure. Substantial τ^2 naturally led to wider CrIs compared to small τ^2 without affecting the point estimate. With moderate inconsistency, the posterior distribution of log ORs deviated from the simulated distribution in all prior structures of log IMOR.

Posterior distribution of τ^2

Posterior median of τ^2 was close to zero in all prior structures of log IMOR for low inconsistency and small τ^2, whereas, as expected, it increased for moderate inconsistency and/or substantial τ^2. For moderate MOD and low inconsistency, posterior median and CrI

for τ^2 were quite similar across all prior structures of log IMOR, whereas for large MOD, posterior median for τ^2 increased slightly with wider widths of CrIs that slightly differed for different assumptions of log IMOR within the hierarchical and identical structure (Figure 7.5). Identical structure led systematically to slightly wider CrIs in most prior structures of log IMOR as compared to hierarchical structure. In addition, the point estimates were slightly larger for identical structure, particularly, for moderate inconsistency and large MOD.

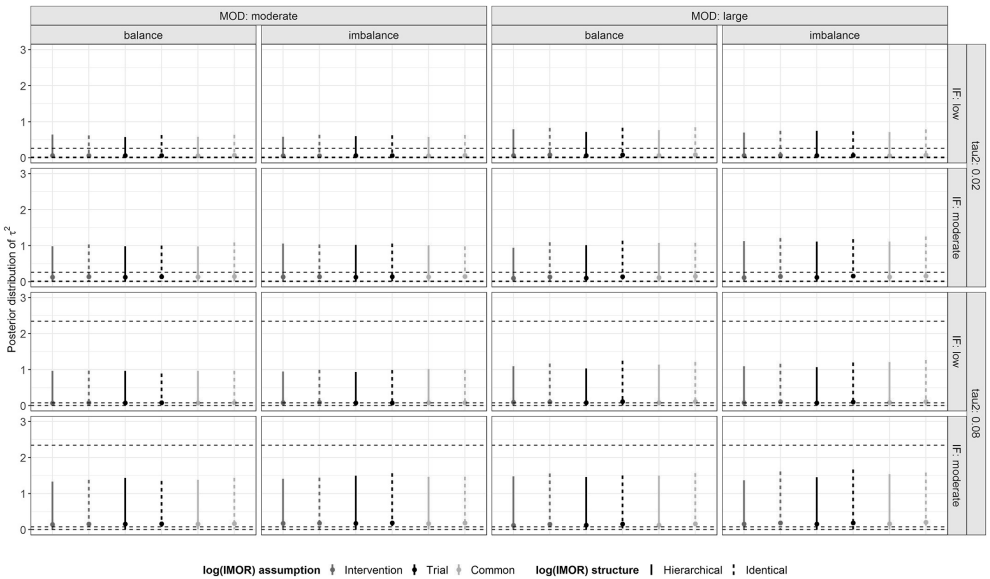

Figure 7.5: Posterior distribution of τ^2 under informative missingness while using pattern-mixture model and accounting for the extent of missing outcome data (moderate, large), balance of missing outcome data (balance, imbalance), extent of τ^2 (small, substantial), and extent of inconsistency (low, moderate). The horizontal dotted lines reflect the 95% interval and median of the simulated distribution of small and substantial τ^2. IF, inconsistency factor; MOD, missing outcome data.

Posterior distribution of IF

Under low inconsistency, the posterior mean of IF was almost zero (ie, evidence of consistency on average) in all prior structures of log IMOR and for all scenarios (Figure 7.6). Overall, CrIs were similarly wider in the presence of substantial τ^2. In the presence of moderate inconsistency, all prior structures of IMOR estimated the true IF, and hence, the point estimates deviated from zero irrespective of extent and balance of MOD.

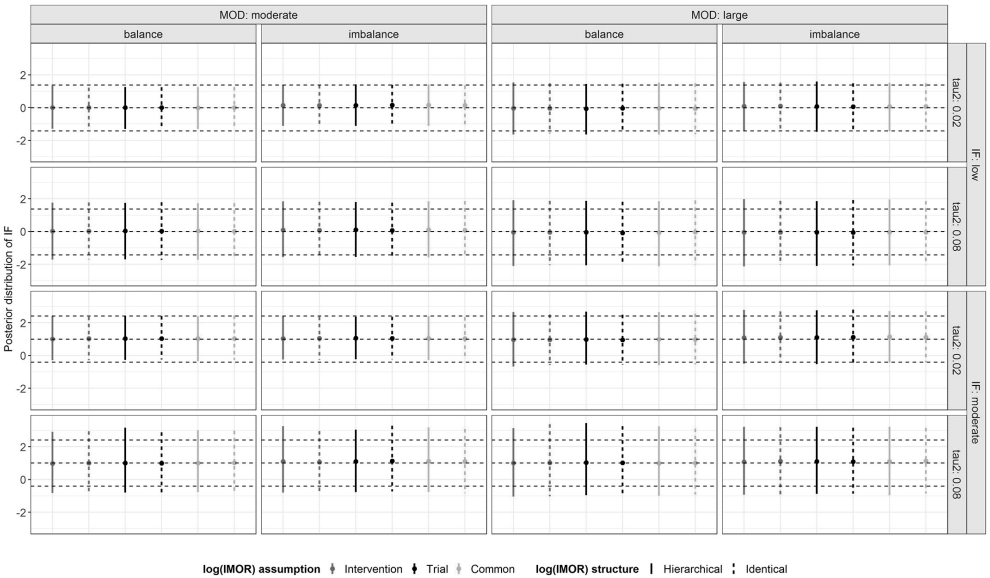

Figure 7.6: Posterior distribution of inconsistency factor (IF) under informative missingness while using pattern-mixture model and accounting for the extent of missing outcome data (moderate, large), balance of missing outcome data (balance, imbalance), extent of τ^2 (small, substantial) and extent of inconsistency (low, moderate). The horizontal dotted lines reflect the 95% interval and mean of the simulated distribution of low and moderate IF. MOD, missing outcome data.

Posterior distribution of probability of being best

The posterior mean of the probability of being best for new intervention was consistently below the simulated values for all prior structures of log IMOR, especially, for large MOD and low inconsistency (Supporting Information S.10; Figure S9). Interestingly, contrary to low inconsistency, moderate inconsistency lowered the posterior mean of the probability of being best the least for all prior structures. Within each scenario, the posterior mean of the probability of being best was similar across all prior structures but slightly larger for unbalanced MOD. Nevertheless, intervention-specific log IMORs led to slightly smaller posterior mean of the probability of being best, especially, for large MOD, moderate inconsistency and small τ^2. The posterior mean of the probability of being best almost overlapped with the simulated values for moderate MOD, small τ^2, and moderate inconsistency. Results on the posterior mean of the probability of being best for old

intervention and placebo can be found in the Supporting Information (results not shown). Overall, different scenarios and prior structures of log IMOR did not impact on the hierarchy of the interventions.

7.6 Discussion

Using three published networks with different extent of MOD as motivating examples, we compared pattern-mixture with selection model while considering six different prior structures of log IMOR that reflected our prior beliefs about the (dis)similarity of log IMORs within the network. Then, on the basis of the results from the motivating examples, we set up a simulation study using empirical-based scenarios to evaluate more in-depth the performance of these prior structures of log IMOR in terms of posterior distribution of log OR, τ^2, IF and probability of being best per intervention. We focused on the performance of prior structures when informative MOD (the most plausible scenario in a medical setting) were analysed under MAR (the recommended primary analysis for MOD). To our knowledge, this is the first simulation study that evaluates statistical modeling of aggregated MOD using Bayesian approaches.

Ultimate goal of the present study was to supplement our observations from our empirical study on these modelling strategies [290]. In our empirical study [290], we used Bland-Altman plots to investigate the degree of agreement among these strategies in terms of NMA estimates. The majority of the networks considered had either low or moderate and balanced MOD. Therefore, we were not able to conclude on the agreement of the strategies when MOD were large or moderate and unbalanced. Furthermore, with an empirical study, we cannot infer on performance measures, such as bias. Consequently, the present simulation study addressed the aforementioned limitations and, additionally, allowed us to investigate the performance of the strategies under different scenarios for the NMA estimates in order to understand the circumstances that may compromise the performance of the strategies.

The last two motivating examples agreed with our empirical study [290], which indicated that, for moderate and large MOD, (hierarchical and identical) intervention-specific prior structure of log IMOR led to larger posterior standard deviation of log ORs as compared to trial-specific and common-within-network prior structures – the latter two led overall to similar posterior

distributions of log ORs. White et al also noticed that the uncertainty around meta-analysis log OR was larger for intervention-specific prior structure while similar for trial-specific and common-within-network prior structures [284]. Our simulation revealed this pattern for large MOD only, regardless of balance of MOD. This performance of intervention-specific prior structure was anticipated as it assumes MOD to be differently informative in different interventions, and therefore, it substantially down-weights trials with moderate or large MOD leading to larger posterior standard deviation of summary log OR [296].

Furthermore, both the present study and our empirical study [290] demonstrated that hierarchical prior structure of log IMOR led to slightly more precise τ^2 compared to identical prior structure, particularly for moderate, unbalanced MOD (Section 7.3). According to our simulation study, this performance was more profound for large MOD with concurrence of inconsistent evidence and/or substantial τ^2. The extent of informative missingness (as quantified via log IMOR) was simulated to vary across the included trials for the same intervention (Equation (7.1) in Supporting Information S.8); however, the identical structure did not capture this variability yielding spuriously narrower CrIs for the study-specific log ORs as compared to hierarchical structure which, in turn, led to relatively larger τ^2 and uncertainty thereof.

The third motivating example indicated that common-within-network prior structure provided slightly more precise estimation of τ^2 compared to intervention- and trial-specific prior structures. Nevertheless, in the simulation study, this pattern was less obvious for moderate, unbalanced MOD, and small τ^2. Possible explanation may be that the motivating example had almost three times more trials than the simulated networks, and in conjunction with the common-within-network being the least data demanding structure of log IMOR, τ^2 was estimated with relatively more precision for this prior structure in the motivating example.

We found that pattern-mixture and selection models gave almost identical results for each prior structure of log IMOR in the motivating examples and simulation study (Supporting Information S.12). While these two models lead to fundamentally opposite factorizations of the joint distribution of the missingness indicator and outcome, the parameter of interest $p^l_{i,k}$ is not affected by this factorization, because, in both models, $p_{i,k}$ is function of $q^l_{i,k}$ and $\phi^l_{i,k}$

(see Equations (7.2) and (7.3)) with the same informative prior distribution being assigned on $\phi^l_{i,k}$. Where these models differ is on the conditional probabilities that define $\phi^l_{i,k}$ (Section 7.2 "Informative missingness odds ratio as missingness parameter"). Nevertheless, if one is interested in investigating the interventions to subgroups of trials that are believed to have different measurement patterns, then pattern-mixture model may be the proper option [283].

For example, (as judged, for instance, by the Cochrane's risk of bias tool; chapter 8 in the work of Higgins and Green [117]), if poorly conducted trials have more MOD than well-conducted trials – and the researcher believes that compared to those leaving poorly conducted trials, patients completing these trials may be more likely to have experienced the beneficial outcome – the researcher should investigate whether the pattern of outcomes in these two trial settings may affect differently the interventions compared. To our knowledge, pattern-mixture model has not been applied yet in series of trials with the aim to provide further insights on the effectiveness of the interventions on subgroups of different patterns of outcome. Instead, if one is interested in the effectiveness of the interventions in the whole population, then pattern-mixture and selection models may be used interchangeably in the analysis of series of trials – although, in principle, the latter is a more natural option [283] as it directly reflects the taxonomy of missingness mechanisms as described by Little and Rubin [280].

Deciding on the assumption for log IMOR shall be primarily tailored to empirical knowledge about the intervention and trial characteristics for the condition under investigation [273, 285]. For example, contrary to active-controlled trials in schizophrenia, placebo-controlled trials lead to greater drop-out rate among patients without improvement in their outcomes [301, 302]. Then, the researcher can consider placebo- and active-specific priors on log IMOR and further investigate the sensitivity of results to using identical and hierarchical structures. In another example, multi-center trials in psychiatry tend to have higher drop-out rate (and hence log IMOR in these trials is more likely to be different from 0) than single-center trials; if log IMORs are believed not to differ among the compared interventions, and the researcher has collected for each trial information on the number of centers, then he/she should assign hierarchical, multi-center-specific, and single-center-specific priors on log IMOR so that log IMORs are different yet related in the corresponding trials. In our simulation study, the proper prior structure of log IMOR was intervention-specific because we assumed placebo to

trigger different missingness mechanisms as opposed to new and old intervention. However, by misspecifying the prior structure using trial-specific or common-within-network prior structure appeared to affect the uncertainty around the log OR leading to narrower CrIs of log OR when MOD were moderate or large. While the inferences about the relative effectiveness of the interventions were not affected in our simulations, the robustness of the inferences for dabigatran 110 mg and rivaroxaban against placebo (third motivating example) were sensible to the prior structure of log IMOR.

In the present study, we addressed aggregated MOD using two popular models for MOD and six different prior structures of log IMOR without accounting, in addition, for important effect modifiers. Van Buuren et al [303] developed a multiple imputation (MI) model that incorporates a delta parameter like IMOR under pattern-mixture model to investigate the degree of departure from MAR in survival analysis in a clinical trial. Extending this model to operate in a collection of trials investigating two or more interventions is an interesting yet unexplored area (to our knowledge) for further work. Provided that we had access to individual patient data (IPD) and enough studies in the network to allow for effect-modification adjustments, MI based on missing not at random (MNAR) assumptions would be a more elegant modelling strategy – though computationally more intensive. This is because MI is already increasingly used for offering a relatively simply and attractive way to account also for the uncertainty induced by imputations (commonly applied under MAR) while adjusting the model for important predictors. In addition, IPD has been often considered as gold standard for synthesizing series of trials as it allows a more rigorous investigation of statistical heterogeneity that – contrary to standard aggregated analysis – protects against the risk for ecological bias, particularly for subject-level characteristics [160]. Since addressing MOD is based on untestable assumptions about missing outcomes (the popular MAR assumption cannot be tested from the observed outcomes), extending 'standard' MI to investigate the sensitivity to MAR via MNAR models offers more flexibility.

The limitations of our study pertain mostly to the simulation setup. Firstly, we used Bayesian approaches as we intended to compare different Bayesian modelling strategies for binary MOD in terms of NMA estimates. Consequently, we preferred not to infer on the performance of the models in terms of frequentist measures, such as type I error, efficiency, and coverage;

contrariwise, our inferences stemmed from the posterior distribution of the NMA estimates for different scenarios and models. Secondly, we considered a simple network of three interventions and two-arm trials with binary outcome data (the most prevalent outcome type in systematic reviews [287]). A more complex network with the addition of multi-arm trials – a 'typical' network in practice [116] – will shed more light on the implications of network complexity on the NMA estimates across different prior structures of log IMOR. For instance, in a complex yet sparse network (where the number of trials and observed comparisons are limited), identical prior structure may perform better to hierarchical structure as it is the least data demanding (alike the common-within-network prior structure). Thirdly, we did not investigate the impact of event frequency since we considered only frequent events. As noted in the work of Carpenter and Kenward, [304] 'if an event (eg, death or a serious side effect) is rare, missing (outcome) data on very few patients can markedly alter estimated event rates,' and therefore, affect substantially the NMA estimates. Fourthly, the degree of unbalanced MOD considered in the simulation setup was much smaller than the total extent of MOD in each trial (Supporting Information S.7).

Consequently, the width of CrI for log OR under common-within-network and trial-specific prior structures (they assume MOD to be equally informative in the whole network and within each trial, respectively, and hence, they down-weight trials with unbalanced MOD in the compared arms [296]) remained narrower than the width of CrI for log OR under intervention-specific prior structure when MOD were unbalanced. A much larger imbalance of MOD may have resulted in more imprecise log OR under common-within-network and trial-specific prior structures. However, we did not observe such extent of imbalance in our empirical study (Supporting Information S.5; Figure S8(f)). Lastly, we dealt with convergence issues (via inspection of the trace and autocorrelation plots) after applying identical common-within-network in both pattern-mixture and selection models; this issue was not tackled after we increased thinning at 6 and 10 (Figures not shown).

Recommendations for the reviewer

- The reviewer should decide in advance on the proper prior structure of log IMOR to address aggregated MOD that best aligns with the condition investigated and the

interventions forming the network; otherwise, misspecification of the prior structure may lead to spurious estimation of the uncertainty around log OR with implications for the conclusions—as shown in the motivating examples and simulation study.

- Pattern-mixture and selection models can be applied interchangeably to infer on the effectiveness of the compared interventions on the whole population.
- Both identical and hierarchical structure may be considered in the context of a sensitivity analysis; though, we expect log IMORs to be different (since the extent of MOD will differ across trial-arms, among other reasons) yet related to each other, and hence, we regard hierarchical structure to be more plausible in practice.

7.7 Conclusions

Assuming MAR on average as a starting point to analyse informative MOD under different prior structures of log IMOR appeared to implicate mainly the precision of the NMA estimates without affecting our conclusions about the effectiveness and the hierarchy of the interventions. Nevertheless, the inferences from the present simulation study were greatly restricted by the scenarios considered. Reviewers should decide already at the protocol of the systematic review on the prior structure of log IMOR according to the condition and interventions investigated. Our results may be also generalized to conventional meta-analyses with binary outcome.

Chapter 7

Supplementary material and coloured figures $1 - 6$ can be found at https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8207

Chapter 8

General Discussion

Clinical trials are not isolated experiments, therefore, their synthesis via meta- analysis has naturally become common and highly valued practice. In rare diseases, due to the small number of available patients, individual trials are often indecisive and, in some cases, contradictory. Thus, for a rare disease, the synthesis of these few clinical trials as part of an exploratory step or a formal drug development plan becomes highly relevant. In such a synthesis, between-study variability (heterogeneity) may be introduced even between two similar trials, due to design, methods and other types of discordances. Accounting for heterogeneity between clinical trials in rare diseases becomes almost infeasible due to the limited number of conducted trials. In order to properly estimate and account for this additional variability, using only a handful of available patient data, robust statistical approaches need to be identified and explored.

In this thesis, I developed and evaluated statistical approaches for evidence synthesis in rare diseases under both a frequentist and a Bayesian framework. Efficiency issues are common under both frameworks within the meta-analytical context of small populations. While the Bayesian framework is often debated, it has been shown to produce robust evidence during the synthesis of summary measures from a few small available trials [12, 13, 178].

Evidence synthesis in small populations

The frequentist and Bayesian scoping literature reviews of the thesis were initially designed to summarize and illustrate available methodologies in evidence synthesis of small populations. The motivation for summarizing evidence-synthesis methods originated from the suboptimal characteristics of a meta-analysis in the presence of a few small (heterogeneous) trials, also shown in Gonnermann et al [11]. Unfortunately, after conducting preliminary searches for both reviews, within the year 2014, only a handful of fitting methodological articles were identified [158, 159].

More specifically, by examining the included methodological articles alongside their titles, one may observe that key terms such as ”rare diseases“, ”small populations“, ”few small trials“ only appeared in the literature after the year 2014; up until then, most meta-analytical, useful for small populations, approaches were either non-existing or they were based on already existing approximate approaches, which were being re-purposed for the design of a new trial

[305, 306]. The three European research projects (ASTERIX, IDeAl and InSPiRe) redirected the focus on small populations' methodology research. Since then, a number of articles have been published, evaluating methods for a meta-analysis of a few and even two small trials [11, 12, 147, 150, 197, 241, 178, 307, 308]. Moreover, even general simulation-based research on evidence synthesis explicitly accounts for very few trials or small population scenarios [309, 310].

Most of the aforementioned recent methodological studies in rare diseases are being conducted through simulation; often authors assume approximate normal effect estimates of dichotomous outcomes on modelling the meta-analysis. Chapter 3 of this thesis conducted a critical evaluation of applied data generating models for a random-effect meta-analysis of dichotomous outcomes in such studies. Even though all evaluated data generating mechanisms were designed to produce the same overall effect, the properties of resulting joint empirical distributions heavily differed, making direct comparisons between methodological studies unreliable [241]. Such concerns extend beyond the small population context, hold broadly for multilevel binomial data studies and may become even more relevant for sparse-event settings (Chapter 4 and 5). The algorithms presented in Chapter 3 can be further generalized to multiple outcome meta-analysis [51], while they have already been generalized in network meta-analysis by Seide et al [311] (pCFixed algorithm) and by Chapter 7 of this thesis [312] (pRandom algorithm).

The latter algorithm - pRandom - has already been (directly) applied on another comparison of modelling strategies of missing outcome data under a frequentist network meta-analysis [313]. Together with Chapter 7, they provide an overall evaluation of missing outcome data modelling strategies under a network meta-analysis. Even though the empirical and simulative characteristics of Chapter 7 were based on prevalent diseases, in a rare disease setting, where often sparsely connected networks of trials exist, it is expected that inferences would become even more prior-driven [136, 137, 314].

Evidence synthesis of a few small clinical trials with rare events

Additionally this thesis points out the improper estimation of the between-study variability when zero event arms increase in a meta-analysis, a situation which may lead to increased

or decreased confidence in treatment efficacy. We further focused on the evaluation of heterogeneity in such settings. Three promising heterogeneity estimators were identified under the frequentist framework (Chapter 4). Under a Bayesian framework (Chapter 5) our results contradicted recent publications in Bayesian meta-analysis of small populations. These publications suggested "weakly" informative priors to stabilize inference and produce robust treatment effects intervals in a meta-analysis of a few small trials. Recommendations on priors to be applied in such small population settings should be made, subject to the special clinical and statistical characteristics of a (network) meta-analysis (such as; sparsity of events, sparsity of network connections), if not completely in a case by case basis.

As no head-to-head comparison of synthesis methods with regards to a few small trials with rare events currently exists, I present a brief re-analysis of example (b) of Chapter 5, based on a number of alternative Bayesian and frequentist strategies, which are explored in either Chapters 3, 4 or 5 (Figure 8.1).

If a practitioner chooses to rely on the frequentist framework, then the simple DerSimonian-Laird estimator risks producing a false positive result as it tends to underestimate true heterogeneity [150, 241, 206]. The Sidik-Jonkman estimator would be a more appropriate choice as shown in Chapter 4 [150]. It should be noted that more sophisticated frequentist models were recently shown to perform very poorly in a few trial meta-analysis of sparse events scenarios [310].

If a practitioner relies on the Bayesian framework, priors that place most of the mass in small values of τ but naturally restrict the range to more plausible values should be preferred, for example a $Uniform(-10, 10)$ on the $log(\tau^2)$ scale [178]. The usual half-normal prior on τ was shown to underestimate true heterogeneity in sparse-event settings; and should not be preferred in such a context [178].

Lastly, future methodological studies in a sparse event meta-analysis in small populations setting should also consider how much importance should be placed in the strict control over operational characteristics and false-positive findings. Especially in the context of major adverse events even the smallest and non-significant signal should be explored.

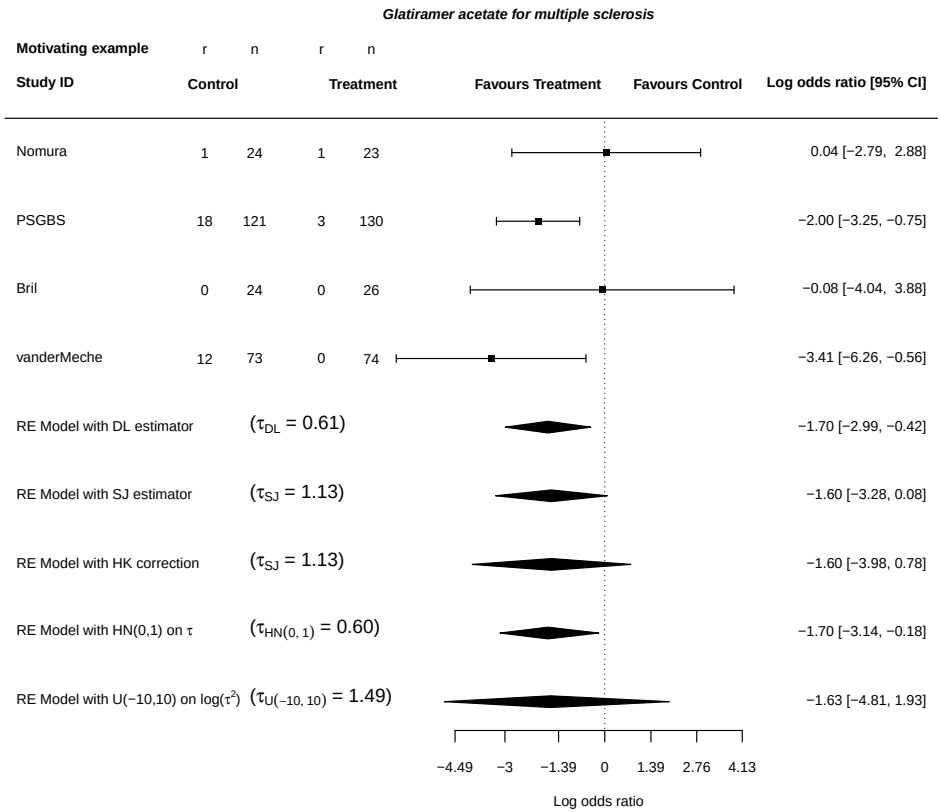

Figure 8.1: Forest plot for the Intravenous immunoglobulin treatment for Guillaine-Barre syndrome [207] reporting the overall effect on endpoint "treatment discontinuation" in terms of log odds ratios. Different modelling options [150, 176, 204, 241, 178] are presented, alongside the resulting trial-specific estimates and their 95% confidence/credible intervals. The 5^{th} row shows the overall effect estimate based on a random-effect model with the DerSimonian Laird estimator. The next lines show a random-effect model with the Sidik and Jonkman estimator and a random effect model with Hartung-Knapp correction, while the last two lines present two binomial-normal hierarchical Bayesian models, one with a half-normal prior on τ and one with a uniform prior on $log(\tau^2)$.

Evidence synthesis accounting for sample-based selection bias

Chapter 6 focused on a pragmatic situation in the evaluation and development of orphan drugs; the informal extrapolation that can occur when results of early phase clinical trials are seen as strengthening the evidence from small Phase III trials. More specifically, the increasing

type I errors and "sample-based" selection biases caused by such an informal combination of exploratory and confirmatory evidence was the main issues addressed. Even though such a combination/extrapolation of historical trial data, which can be considered as part of the design of a new clinical trial, was explored in Chapter 6, the design of a new clinical trial was not directly considered in this thesis.

Except for the above setting, this "sample-based" selection bias could be of general interest for a meta-analysis of two trials, as the conduct of a succeeding clinical trial is usually conditional on previously observed positive trial results. In such a situation a simple meta-analysis of two (or even three) trials may yield positively biased overall effects and alternative statistical approaches, such as variations of those described in Chapter 6, may be needed.

Conclusion

Meta-analysis in small populations faces certain issues that derive mostly from the limited number of available and often heterogeneous small trials to be synthesized. When such an analysis is conducted within a small population context, it should be applied with caution and even further tailored to account for specific small population characteristics, a few of which were extensively discussed in this thesis.

Appendix

List of Abbreviations

AD: Aggregate study-level data

BDR: Bayesian double-regression

BFDR: Bayesian flexible double-regression

CC: Continuity correction

CI: Confidence Interval

CMA: Cumulative meta-analysis

CrI: Credible Interval

DGM: Data generating model

DR: Double-regression

DRC: Corrected double-regression

EMA: European Medicines Agency

EU-FP7: European Union - Framework project 7

FDA: Federal Drug Agency

FE: Fixed-effect

GBS: Guillain-Barre syndrome

GL-3: globotriaosylceramide

IMOR: Informative missingness odds ratio

HK: Hartung and Knapp

IPD: Individual patient data

IQR: Interquartile range

IVIG: Intravenous immunoglobulin

LOR: Log odds ratio

MA: Meta-analysis

MAR: Missing at random

MCMC: Markov chain Monte Carlo

MH: Mantel-Haenszel

MMA: Multiple outcome meta-analysis / multivariate meta-analysis

MOD: Missing outcome data

MSE: Mean square error

NMA/MTC: Network meta-analysis / mixed treatments meta-analysis

OR: Odd-ratio

PA: pAverage

PCF: pCFixed

PH: Proportional hazards

PHK: Phenylketonuria

PMA: Pairwise meta-analysis

PR: pRandom

RCT: Randomized controlled trial

RE: Random-effects

REML: Restricted maximum-likelihood

RR: Risk-ratio

SMA: Sequential meta-analysis

SR: Systematic review

TSA: Trial sequential meta-analysis

Appendix

Heterogeneity estimators

dl: DerSimonian-Laird

dlp: Positive DerSimonian Laird

dl2: Two-step Der Simonian Laird

he: Hedges

he2: Two step Hedges

hm: Hartung - Makambi

hs: Hunter-Schmidt

ipm: Improved Paul-Mandel

ml: Maximum Likelihood

mvvc: Model error variance - vc

pm: Paul-Mandel

rb0: Rukhin Bayes zero estimator

rbp: Rukhin Bayesian positive

reml: Restricted Maximum likelihood

sj: Positive Sidik-Jonkman

List of Tables

Appendix

List of Figures

Appendix

Bibliography

Bibliography

[1] Charlotte Rodwell and Ségolène Aymé. Rare disease policies to improve care for patients in Europe. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1852(10):2329–2335, 2015.

[2] Committee for medicinal products hor human use (CHMP). *Guideline on clinical trials in small populations*. European Medicines Agency, 2006.

[3] Joshua J Gagne, Lauren Thompson, Kelly O'Keefe, and Aaron S Kesselheim. Innovative research methods for studying treatments for rare diseases: methodological review. *Bmj*, 349:g6802, 2014.

[4] Lisa V Hampson, John Whitehead, Despina Eleftheriou, and Paul Brogan. Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Statistics in Medicine*, 33(24):4186–4201, 2014.

[5] Samir Gupta, Marie E Faughnan, George A Tomlinson, and Ahmed M Bayoumi. A framework for applying unfamiliar trial designs in studies of rare diseases. *Journal of Clinical Epidemiology*, 64(10):1085–1094, 2011.

[6] ASTERIX project, FP7-HEALTH, Grant no 603160.

[7] IDEAL project, FP7-HEALTH, Grant no 602552.

[8] InSPiRe project, FP7-HEALTH, Grant no 602144.

[9] Michael Borenstein, Larry V Hedges, Julian Higgins, and Hannah R Rothstein. Introduction to meta-analysis. 2009.

[10] Georgia Salanti. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods*, 3(2):80–97, 2012.

[11] Andrea Gonnermann, Theodor Framke, Anika Großhennig, and Armin Koch. No solution yet for combining two independent studies in the presence of heterogeneity. *Statistics in Medicine*, 34(16):2476–2480, 2015.

[12] Kristina Weber, Rob Hemmings, and Armin Koch. How to use prior knowledge and still give new data a chance? *Pharmaceutical statistics*, 17(4):329–341, 2018.

[13] Tim Friede, Christian Röver, Simon Wandel, and Beat Neuenschwander. Meta-analysis of few small studies in orphan diseases. *Research Synthesis Methods*, 8(1):79–91, 2017.

[14] CBER CDER. *Rare Diseases: Common Issues in Drug Development, Guidance for Industry*. FDA, 2015.

[15] FDA. Guidance for the use of Bayesian statistics in medical device clinical trials. *Guidance for Industry and FDA staff*, pages 1–50, 2010.

[16] Joanna IntHout, John P A Ioannidis, and George F Borm. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC medical research methodology*, 14(1):25, 2014.

[17] Christian Röver, Guido Knapp, and Tim Friede. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology*, 15(1):99, 2015.

[18] Daniel J O Connor, Robert James Hemmings, Daniel J O Connor, and Robert James Hemmings. Expert Opinion on Orphan Drugs Coping with small populations of patients in clinical trials Coping with small populations of patients in clinical trials. *Expert Opinion on Orphan Drugs*, 2(8):765–768, 2014.

[19] European Medicines Agency - Guidelines on Clinical Trials in Small Populations, 2006.

[20] Catherine Cornu, Behrouz Kassai, Roland Fisch, Catherine Chiron, Corinne Alberti, and Renzo Guerrini. Experimental designs for small randomised clinical trials : an algorithm for choice. *Orphanet Journal of Rare Diseases*, 8(48):1–12, 2013.

[21] Fabio Aiello, Massimo Attanasio, and Fabio Tinè. Assessing covariate imbalance in meta-analysis studies. *Statistics in medicine*, 30(22):2671–82, 2011.

[22] Jos Verbeek, Jani Ruotsalainen, and Jan L Hoving. Synthesizing study results in a systematic review. *Scandinavian journal of work, environment & health*, 38(3):282–90, 2012.

[23] Sander Greenland. Accounting for uncertainty about investigator bias: disclosure is informative. *Journal of epidemiology and community health*, 63(8):593–8, 2009.

[24] Gregory Camilli, Jimmy de la Torre, and Chia-Yi Chiu. A noncentral t regression model for meta-analysis. *Journal of Educational and Behavioral Statistics*, 35(2):125–153, 2010.

[25] National Research Council. Combining information: Statistical issues and opportunities for research., 1992.

[26] Julian P T Higgins, Simon G Thompson, and David J Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159, 2009.

[27] Yeojin Chung, Sophia Rabe-Hesketh, and In-Hee Choi. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in medicine*, 32(23):4071–89, 2013.

[28] Eugene Demidenko, James Sargent, and Tracy Onega. Random effects coefficient of determination for mixed and meta-analysis models. *Communications in statistics: theory and methods*, 41(6):953–969, 2012.

[29] Michael J Malloy, Luke A Prendergast, and Robert G Staudte. Transforming the Model T: random effects meta-analysis with stable weights. *Statistics in medicine*, 32(11):1842–64, 2013.

[30] Wolfgang Viechtbauer. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical software*, 36(3):48, 2010.

[31] Alexander J Sutton, Nicola J Cooper, and David R Jones. Evidence synthesis as the key to more coherent and efficient research. *BMC medical research methodology*, 9(29):1–14, 2009.

[32] Alison C Goudie, Alexander J Sutton, David R Jones, and Alison Donald. Empirical assessment suggests that existing evidence could be used more fully in designing randomized controlled trials. *Journal of clinical epidemiology*, 63(9):983–91, 2010.

[33] John Ioannidis and Fotini Karassa. The need to consider the wider agenda in systematic reviews and meta-analyses. *BMJ: British Medical Journal*, 341:c4875:762–765, 2010.

[34] Michael Rotondi and Allan Donner. Sample size estimation in cluster randomized trials: An evidence-based perspective. *Computational Statistics & Data Analysis*, 56(5):1174–1187, 2012.

[35] Julian P T Higgins and Simon G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002.

[36] Jørn Wetterslev, Kristian Thorlund, Jesper Brok, and Christian Gluud. Estimating required information size by quantifying diversity in random-effects model meta-analyses. *BMC medical research methodology*, 9(86):12, 2009.

[37] Rebecca J. Hardy and Simon G. Thompson. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, 15(6):619–629, 1996.

[38] Rebecca M Turner, David J Spiegelhalter, Gordon C S Smith, and Simon G Thompson. Bias modelling in evidence synthesis. *Journal of Royal Statistical Society (A)*, 172(1):21–47, 2014.

[39] S. Ahn and B. J. Becker. Incorporating Quality Scores in Meta-Analysis. *Journal of Educational and Behavioral Statistics*, 36(5):555–585, 2011.

[40] Ying Yuan and Roderick J A Little. Meta-analysis of studies with missing data. *Biometrics*, 65(2):487–96, 2009.

[41] Gerta Rücker, Guido Schwarzer, James R Carpenter, Harald Binder, and Martin Schumacher. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics (Oxford, England)*, 12(1):122–42, 2011.

[42] Theo Stijnen, Taye H Hamza, and Pinar Ozdemir. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine*, 29(29):3046–3067, 2010.

[43] Peter W Lane. Meta-analysis of incidence of rare events. *Statistical methods in medical research*, 22(2):117–132, 2013.

[44] Christy Chuang-Stein and Mohan Beltangady. Reporting cumulative proportion of subjects with an adverse event based on data from multiple studies. *Pharmaceutical statistics*, 10(1):3–7, 2011.

[45] Susan Gruber and Mark J van der Laan. An application of targeted maximum likelihood estimation to the meta-analysis of safety data. *Biometrics*, 69(1):254–62, 2013.

[46] Monica M Bennett, Brenda J Crowe, Karen L Price, James D Stamey, and John W. Seaman. Comparison of Bayesian and Frequentist Meta-Analytical Approaches for Analyzing Time to Event Data. *Journal of Biopharmaceutical Statistics*, 23(1):129–145, 2013.

[47] Duncan Chambers, Mark Rodgers, and Nerys Woolacott. Not only randomized controlled trials, but also case series should be considered in systematic reviews of rapidly developing technologies. *Journal of clinical epidemiology*, 62(12):1253–1260, 2009.

[48] Siew Wan Hee and Nigel Stallard. Designing a series of decision-theoretic phase II trials in a small population. *Statistics in medicine*, 31(30):4337–51, 2012.

[49] Marie-Cécile Le Deley, Karla V Ballman, Julien Marandet, and Daniel Sargent. Taking the long view: how to design a series of Phase III trials to maximize cumulative therapeutic benefit. *Clinical trials (London, England)*, 9(3):283–92, 2012.

[50] Richard Sposto and Daniel O Stram. A strategic view of Randomized Trial design in low-incidence paediatric cancer. *Statistics in medicine*, (18):1183–1197, 1999.

[51] Richard D . Riley. Multivariate Meta-Analysis : The Effect of Ignoring Within-Study Correlation. *Journal of Royal Statistical Society (A)*, 172(4):789–811, 2014.

[52] Richard D Riley, John R Thompson, and Keith R Abrams. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, 9(1):172–186, 2008.

[53] Dan Jackson, Richard Riley, and Ian R White. Multivariate meta-analysis: potential and promise. *Statistics in medicine*, 30(20):2481–98, 2011.

[54] Larry V Hedges. Comment on ' Multivariate meta-analysis : Potential and promise '. *Statistics in medicine*, page 1, 2011.

[55] Roger M Harbord. Commentary on ' Multivariate meta-analysis : potential and promise '. *Statistics in medicine*, page 2, 2011.

[56] Antonio Gasparrini, Ben Armstrong, and Mike G Kenward. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in medicine*, 31(29):3821–3839, 2012.

[57] J Martin Bland. Comments on ' Multivariate meta-analysis : Potential and promise ' by Jackson et al . , Statistics in Medicine. *Statistics in medicine*, page 2, 2011.

[58] David R Cox. Multivariate meta-analysis : A comment. *Statistics in medicine*, page 2, 2011.

[59] George F Borm and A Rogier T Donders. Updating meta-analyses leads to larger type I errors than publication bias. *Journal of clinical epidemiology*, 62(8):825–830, 2009.

[60] Jesper Brok, Kristian Thorlund, Jørn Wetterslev, and Christian Gluud. Apparently conclusive meta-analyses may be inconclusive–Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. *International journal of epidemiology*, 38(1):287–98, 2009.

[61] Ingeborg van der Tweel and Casper Bollen. Sequential meta-analysis: an efficient decision-making tool. *Clinical trials*, 7(2):136–46, 2010.

[62] Branko Miladinovic, Ambuj Kumar, Iztok Hozo, Helen Mahony, and Benjamin Djulbegovic. Trial sequential analysis may be insufficient to draw firm conclusions regarding statistically significant treatment differences using observed intervention effects: a case study of meta-analyses of multiple myeloma trials. *Contemporary clinical trials*, 34(2):257–61, 2013.

[63] Branko Miladinovic, Rahul Mhaskar, Iztok Hozo, Ambuj Kumar, Helen Mahony, and Benjamin Djulbegovic. Optimal information size in trial sequential analysis of time-to-event outcomes reveals potentially inconclusive results because of the risk of random error. *Journal of clinical epidemiology*, 66(6):654–9, 2013.

[64] Pantelis G Bagos and Georgios K Nikolopoulos. Generalized least squares for assessing trends in cumulative meta-analysis with applications in genetic epidemiology. *Journal of clinical epidemiology*, 62(10):1037–44, 2009.

[65] Peter Herbison, Jean Hay-Smith, and William J Gillespie. Meta-analyses of small numbers of trials often agree with longer-term results. *Journal of clinical epidemiology*, 64(2):145–53, 2011.

[66] Tiago V Pereira and John P A Ioannidis. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of Clinical Epidemiology*, 64(10):1060–1069, 2011.

[67] Kristian Thorlund, P J Devereaux, Jørn Wetterslev, Gordon Guyatt, John P A Ioannidis, Lehana Thabane, Lise-Lotte Gluud, Bodil Als-Nielsen, and Christian Gluud. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *International journal of epidemiology*, 38(1):276–86, 2009.

[68] Eveline Nüesch and Peter Jüni. Commentary: Which meta-analyses are conclusive? *International journal of epidemiology*, 38(1):298–303, 2009.

[69] Georgina Imberger, Jørn Wetterslev, and Chistian Gluud. Trial sequential analysis has the potential to improve the reliability of conclusions in meta-analysis. *Contemporary clinical trials*, 36(1):254–5, 2013.

[70] Branko Miladinovic, Iztok Hozo, and Benjamin Djulbegovic. Trial sequential boundaries for cumulative meta-analyses. *The Stata Journal*, 13(1):77–91, 2013.

[71] John Whitehead. *The design and analysis of sequential clinical trials*. John Wiley & Sons, 1997.

[72] Kristian Thorlund, Georgina Imberger, Jørn Wetterslev, Jesper Brok, and Christian Gluud. Comments on ' Sequential meta-analysis : an efficient decision-making tool ' by I van der Tweel and C Bollen. *Clinical Trials*, pages 752–753, 2010.

[73] Ingeborg van der Tweel and Casper Bollen. Response to Letter from K Thorlund et al . *Clinical Trials*, page 1, 2010.

[74] Putri W Novianti, Kit C B Roes, and Ingeborg van der Tweel. Estimation of between-trial variance in sequential meta-analyses: a simulation study. *Contemporary clinical trials*, 37(1):129–38, 2014.

[75] David K Turok, Eve Espey, Alison B Edelman, Pamela S Lotke, Eva H Lathrop, Stephanie B Teal, Janet C Jacobson, Sara E Simonsen, and Kenneth F Schulz. The methodology for developing a prospective meta-analysis in the family planning community. *Trials*, 12(1):104, 2011.

[76] Julian P T Higgins, Anne Whitehead, and Mark Simmonds. Sequential methods for random-effects meta-analysis. *Statistics in medicine*, 30(9):903–21, 2011.

[77] Georgina Imberger, Christian Gluud, and Jørn Wetterslev. Comments on 'Sequential methods for random-effects meta-analysis' by J. P. Higgins, A. Whitehead and M. Simmonds. *Statistics in medicine*, 30(24):2965–6, 2011.

[78] Jonathan J Shuster and Josef Neu. A Pocock approach to sequential meta-analysis of clinical trials. *Research synthesis methods*, 4(3), 2013.

[79] Lisa M Askie, Louise A Baur, Karen Campbell, Lynne A Daniels, Kylie Hesketh, Anthea Magarey, Seema Mihrshahi, Chris Rissel, John Simes, Barry Taylor, et al. The early prevention of obesity in children (epoch) collaboration-an individual patient data prospective meta-analysis. *BMC public health*, 10(1):728, 2010.

[80] John P A Ioannidis. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 181(8):488–93, 2009.

[81] Kristian Thorlund and Edward J Mills. Sample size and power considerations in network meta-analysis. *Systematic reviews*, 1(1):41, 2012.

[82] Kristian Thorlund and Edward Mills. Stability of additive treatment effects in multiple treatment comparison meta-analysis: A simulation study. *Clinical Epidemiology*, 4(1):75–85, 2012.

[83] Huseyin Naci and Alec B O'Connor. Assessing comparative effectiveness of new drugs before approval using prospective network meta-analyses. *Journal of clinical epidemiology*, 66(8):812–6, 2013.

[84] Aïda Bafeta, Ludovic Trinquart, Raphaèle Seror, and Philippe Ravaud. Analysis of the systematic reviews process in reports of network meta-analyses: methodological systematic review. *BMJ (Clinical research ed.)*, 347:1–12, 2013.

[85] Richard D Riley, Paul C Lambert, and Ghada Abo-Zaid. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ (Clinical research ed.)*, 340:c221(7745), 2010.

[86] Stephanie A Kovalchik. Aggregate-data estimation of an individual patient data linear random effects meta-analysis with a patient covariate-treatment interaction term. *Biostatistics (Oxford, England)*, 14(2):273–83, 2013.

[87] Richard J Lilford, A Richardson, A Stevens, R Fitzpatrick, S Edwards, F Rock, and JL Hutton. Issues in methodological research: perspectives from researchers and commissioners. *Health technology assessment (Winchester, England)*, 5(8):1–57, 2001.

[88] Matthias Egger and Davey G Smith. Misleading meta-analysis. *BMJ: British Medical Journal*, 311(7007):753, 1995.

[89] Joanna IntHout, John PA Ioannidis, and George F Borm. Obtaining evidence by a single well-powered trial or several modestly powered trials. *Statistical methods in medical research*, 25(2):538–552, 2016.

[90] TC Collaboration. Review manager (revman). *Copenhagen: The Nordic Cochrane Centre*, 2008.

[91] Michael Borenstein, Larry Hedges, Julian Higgins, and Hannah Rothstein. Comprehensive meta-analysis version 2. pages 1–104, 2005, Website: https://www.meta-analysis.com/downloads/Meta-Analysis-Manual.pdf, Accessed: 11/03/2020.

[92] Rebecca DerSimonian and Raghu Kacker. Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, 28(2):105–114, 2007.

[93] Gerta Rücker, Guido Schwarzer, James R Carpenter, and Martin Schumacher. Undue reliance on i 2 in assessing heterogeneity may mislead. *BMC medical research methodology*, 8(1):79, 2008.

[94] Richard D Riley. Multivariate Meta-Analysis : The Effect of Ignoring Within-Study Correlation. *Journal of Royal Statistical Society (A)*, 172(4):789–811, 2009.

[95] Franz Koenig, Jim Slattery, Trish Groves, Thomas Lang, Yoav Benjamini, Simon Day, Peter Bauer, and Martin Posch. Sharing clinical trial data on patient level: opportunities and challenges. *Biometrical Journal*, 57(1):8–26, 2015.

[96] Pablo E Verde and Christian Ohmann. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research synthesis methods*, 6(1):45–62, 2015.

[97] Robin Harbour and Juliet Miller. A new system for grading recommendations in evidence based guidelines. *BMJ (Clinical research ed.)*, 323(7308):334–336, 2001.

[98] Bruce G Charlton. Medical practice and the double-blind, randomized controlled trial. *The British Journal of General Practice*, 41(350):355–356, 1991.

[99] Sarah JL Edwards, Richard J Lilford, David Braunholtz, and Jennifer Jackson. Why 'underpowered' trials are not necessarily unethical. *Lancet*, 350(9080):804–807, 1997.

[100] Caridad Pontes, Juan M Fontanet, Monica Gomez-Valent, J Rios Guillermo, Roser Vives V, Rosa Morros, Jorge Martinalbo, Josep Torrent-Farnell, and Ferran Torres. Milestones On Orphan Medicinal Products Development: The 100 First Drugs for Rare Diseases Approved Throughout Europe. *Clinical Therapeutics*, 37(8):e132, 2016.

[101] Swati Biswas, Diane D Liu, Jack J Lee, and Donald A Berry. Bayesian Clinical Trials at the University of Texas M.D. Anderson Cancer Center. *Clinical Trials*, 6(3):205–216, 2010.

[102] European Medicines Agency - ICH Topic E9 Statistical Principles for Clinical Trials, 1998.

[103] Lucinda Billingham, Kinga Malottki, and Neil Steven. Small sample sizes in clinical trials: a statistician's perspective. *Clinical Investigation*, 2(7):655–657, 2012.

[104] Lusine Abrahamyan, Ivan R Diamond, Sindhu R Johnson, and Brian M Feldman. A new toolkit for conducting clinical trials in rare disorders. *Journal of Population Therapeutics and Clinical Pharmacology*, 21(1):66–78, 2014.

[105] David A Schoenfeld. Bayesian design using adult data to augment pediatric trials. *Clin Trials*, 6(4):297–304, 2012.

[106] Brian Neelon and a. James O'Malley. Bayesian Analysis Using Power Priors with Application to Pediatric Quality of Care. *Journal of Biometrics & Biostatistics*, 103(1):1–9, 2010.

[107] David L DeMets. Methods for combining randomized clinical trials: strengths and limitations. *Statistics in medicine*, 6(3):341–348, 1987.

[108] European Medicines Agency - Points to consider on application with 1. Meta-analyses; 2. One pivotal study, 1998.

[109] Gene V Glass. Primary, secondary, and meta-analysis of research. *Educational researcher*, 5(10):3–8, 1976.

[110] Lawrence D Cohn and Betsy J Becker. How meta-analysis increases statistical power. *Psychological methods*, 8(3):243–253, 2003.

[111] European Medicines Agency. Mozobil (plerixafor) - Assessment report. Technical report, 2009.

[112] European Medicines Agency. Wakix (pitolisant) - Assessment report. pages 1–91, 2015.

[113] European Medicines Agency. Darzalex (daratumumab) - Assessment report. pages 1–119, 2016.

[114] Georgia Salanti, A E Ades, and John P A Ioannidis. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis : an overview and tutorial. *Journal of Clinical Epidemiology*, 64(2):163–171, 2011.

[115] Dulal K. Bhaumik, Anup Amatya, Sharon-Lise Normand, Joel Greenhouse, Eloise Kaizar, Brian Neelon, and Robert D. Gibbons. Meta-Analysis of Rare Binary Adverse Event Data. *Journal of the American Statistical Association*, 107(498):555–567, 2012.

[116] Adriani Nikolakopoulou, Anna Chaimani, Areti Angeliki Veroniki, Haris S Vasiliadis, Christopher H Schmid, and Georgia Salanti. Characteristics of networks of interventions: a description of a database of 186 published networks. *PloS one*, 9(1), 2014.

[117] Julian Higgins and Sally Green. Cochrane Handbook for Systematic Reviews of Interventions. *The Cochrane Collaboration*, page Version 5.1.0, 201.

[118] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models Author(s):. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

[119] Hans C van Houwelingen, Lidia R Arends, and Theo Stijnen. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in medicine*, 21(4):589–624, 2002.

[120] Yinghui Wei and Julian P T Higgins. Bayesian multivariate meta-analysis with multiple outcomes. *Statistics in Medicine*, 32(17):2911–2934, 2013.

[121] Dan Jackson, Jack Bowden, and Rose Baker. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference*, 140(4):961–970, 2010.

[122] Yong Chen, Sheng Luo, Haitao Chu, Xiao Su, and Lei Nie. An Empirical Bayes Method for Multivariate Meta-analysis with an Application in Clinical Trials. *Communications in statistics: theory and methods*, 43(16):3536–3551, 2014.

[123] Dimitris Mavridis and Georgia Salanti. A practical introduction to multivariate meta-analysis. *Statistical methods in medical research*, 22(2):133–158, 2012.

[124] Sandra M Hurtado Rúa, Madhu Mazumdar, and Robert L Strawderman. The choice of prior distribution for a covariance matrix in multivariate meta-analysis: a simulation study. *Statistics in medicine*, 34(30):4083–4104, 2015.

[125] Danielle L Burke, Sylwia Bujkiewicz, and Richard D Riley. Bayesian bivariate meta-analysis of correlated effects : Impact of the prior distributions on the borrowing of strength , and joint inferences. 27(2):428–450, 2018.

[126] Sylwia Bujkiewicz, John R Thompson, Alex J Sutton, Nicola J Cooper, Mark J Harrison, Deborah P M Symmons, and Keith R Abrams. Multivariate meta-analysis of mixed outcomes : a Bayesian approach. *Statistics in medicine*, 32(22):3926–3943, 2013.

[127] David C Hoaglin, Neil Hawkins, Jeroen P Jansen, David A Scott, Robbin Itzler, Joseph C Cappelleri, Cornelis Boersma, David Thompson, and Kay M Larholt. Conducting Indirect-Treatment-Comparison and Network-Meta-Analysis Studies : Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices : Part 2. *Value in Health*, 14(4):429–437, 2011.

[128] Jeroen P Jansen, Rachael Fleurence, Beth Devine, and Robbin Itzler. Interpreting Indirect Treatment Comparisons and Network Meta-Analysis for Health-Care Decision Making : Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices : Part 1. *JVAL*, 14(4):417–428, 2011.

[129] Orestis Efthimiou, Thomas P A Debray, Gert van Valkenhoef, Sven Trelle, Klea Panayidou, Karel G M Moons, Johannes B Reitsma, Aijing Shang, and Georgia Salanti. GetReal in network meta-analysis: a review of the methodology. *Research Synthesis Methods*, 7(3):236–263, 2016.

[130] Hwanhee Hong, Bradley P Carlin, Tatyana A Shamliyan, Jean F Wyman, Rema Ramakrishnan, Francois Sainfort, and Robert L Kane. Comparing Bayesian and Frequentist Approaches for Multiple Outcome Mixed Treatment Comparisons. *Medical Decision Making*, 33(5):702–714, 2013.

[131] David J Spiegelhalter, Keith R Abrams, and Jonathan P Myles. *Bayesian approaches to clinical trials and health-care evaluation*, volume 3. John Wiley & Sons, 2004.

[132] Teresa C Smith, David J Spiegelhalter, and Andrew Thomas. Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine*, 14(24):2685–2699, 1995.

[133] Guobing Lu and AE Ades. Assessing Evidence Inconsistency in Mixed Treatment Comparisons. *Journal of the American Statistical Association*, 101(474):447–459, 2006.

[134] Guobing Lu and Ae Ades. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics (Oxford, England)*, 10(4):792–805, 2009.

[135] J P Higgins and A Whitehead. Borrowing strength from external trials in a meta-analysis. *Statistics in medicine*, 15(24):2733–2749, 1996.

[136] Rahul Mhaskar, Jasmina Redzepovic, Keith Wheatley, Otavio Augusto Camara Clark, Branko Miladinovic, Axel Glasmacher, Ambuj Kumar, and Benjamin Djulbegovic. Bisphosphonates in multiple myeloma: a network meta-analysis. *The Cochrane database of systematic reviews*, (5), 2012.

[137] Kavi J. Littlewood, Kyoko Higashi, Jeroen P. Jansen, Gorana Capkun-Niggli, Maria Magdalena Balp, Gerd Doering, Harm a W M Tiddens, and Gerhild Angyalosi. A network meta-analysis of the efficacy of inhaled antibiotics for chronic Pseudomonas infections in cystic fibrosis. *Journal of Cystic Fibrosis*, 11(5):419–426, 2012.

[138] S M Goring, P Gustafson, Y Liu, S Saab, S K Cline, and R W Platt. Disconnected by design : analytic approach in treatment networks having no common comparator. 7(4):420–432, 2016.

[139] Fujian Song, Allan Clark, Max O Bachmann, and Jim Maas. Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons. *BMC Medical Research Methodology*, 12:138, 2012.

[140] Teresa Greco, Giovanni Landoni, Giuseppe Biondi-Zoccai, Fabrizio D'Ascenzo, and Alberto Zangrillo. A Bayesian network meta-analysis for binary outcome: how to do it. *Statistical methods in medical research*, 25(5):1757–1773, 2016.

[141] Kevin Carroll and Robert Hemmings. On the need for increased rigour and care in the conduct and interpretation of network meta-analyses in drug development. *Pharmaceutical Statistics*, 15(2):135–142, 2016.

[142] Hong Zhao, James S Hodges, Haijun Ma, Qi Jiang, and Bradley P Carlin. Hierarchical Bayesian approaches for detecting inconsistency in network meta-analysis. *Statistics in Medicine*, 35(20):3524–3536, 2016.

[143] Stephen Senn. Added values controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine*, 23(24):3729–3753, 2004.

[144] Paul C Lambert, Alex J Sutton, Paul R Burton, Keith R Abrams, and David R Jones. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15):2401–2428, 2005.

[145] Olga Gajic-Veljanoski, Angela M Cheung, Ahmed M Bayoumi, and George Tomlinson. The choice of a noninformative prior on between-study variance strongly affects predictions of future treatment effect. *Medical decision making : an international journal of the Society for Medical Decision Making*, 33(3):356–368, 2013.

[146] Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

[147] Tim Friede, Christian Röver, Simon Wandel, and Beat Neuenschwander. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biometrical Journal*, 59(4):658–671, 2017.

[148] Olha Bodnar, Alfred Link, Barbora Arendacká, Antonio Possolo, and Clemens Elster. Bayesian estimation in random effects meta-analysis using a non-informative prior. *Statistics in Medicine*, 36(2):378–399, 2017.

[149] O Kuss. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. *Statistics in medicine*, 34(7):1097–1116, 2015.

[150] Konstantinos Pateras, Stavros Nikolakopoulos, Dimitris Mavridis, and Kit CB Roes. Interval estimation of the overall treatment effect in a meta-analysis of a few small studies with zero events. *Contemporary clinical trials communications*, 9:98–107 X, 2018.

[151] Kirsty M Rhodes, Rebecca M Turner, and Julian P T Higgins. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of clinical epidemiology*, 68(1):52–60, 2015.

[152] Rebecca M Turner, Jonathan Davey, Mike J Clarke, Simon G Thompson, and Julian PT Higgins. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *International journal of epidemiology*, 41(3):818–27, 2012.

[153] Rebecca M Turner, Dan Jackson, Yinghui Wei, Simon G Thompson, and Julian P T Higgins. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in medicine*, 34(6):984–998, 2015.

[154] Eleanor M Pullenayegum. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Statistics in Medicine*, 30(26):3082–3094, 2011.

[155] Kirsty M Rhodes, Rebecca M Turner, Ian R White, Dan Jackson, J Spiegelhalter, and Julian P T Higgins. Implementing informative priors for heterogeneity in meta-analysis using meta-regression and pseudo data. (August), 2016.

[156] Kirsty M Rhodes, M Turner, and Julian P T Higgins. Empirical evidence about inconsistency among studies in a pair-wise meta- analysis. (November 2015), 2016.

[157] Kristian Thorlund, Lehana Thabane, and Edward J Mills. Modelling heterogeneity variances in multiple treatment comparison meta-analysis – Are informative priors the better solution ? *BMC Medical Research Methodology*, 13(2), 2013.

[158] Andrew L Rukhin. Estimating common mean and heterogeneity variance in two study case. *Statistics and Probability Letters*, 82(7):1318–1325, 2012.

[159] Andrew L Rukhin. Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 75(3):451–469, 2013.

[160] Thomas P A Debray, Karel G M Moons, Gert van Valkenhoef, Orestis Efthimiou, Noemi Hummel, Rolf H H Groenwold, and Johannes B Reitsma. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Research Synthesis Methods*, (August), 2015.

[161] T. P. Debray, E. Schuit, O. Efthimiou, J. B. Reitsma, J. P. Ioannidis, G. Salanti, and K. G. Moons. An overview of methods for network meta-analysis using individual participant data: when do benefits arise? *Statistical Methods in Medical Research*, pages 1–14, 2016.

[162] Michael J Crowther, Richard D Riley, Jan A Staessen, Jiguang Wang, Francois Gueyffier, and Paul C Lambert. Individual patient data meta-analysis of survival data using Poisson regression models. *BMC Medical Research Methodology*, 12(1):34, 2012.

[163] a J Sutton, D Kendrick, and C a C Coupland. Meta-analysis of individual- and aggregate-level data. *Statistics in medicine*, 27(5):651–69, 2008.

[164] Sarah Donegan, Paula Williamson, Umberto D'Alessandro, and Catrin Tudur Smith. Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: individual patient-level covariates versus aggregate trial-level covariates. *Statistics in medicine*, 31(29):3840–57, 2012.

[165] Pedro Saramago, Alex J Sutton, Nicola J Cooper, and Andrea Manca. Mixed treatment comparisons using aggregate and individual participant level data. *Statistics in medicine*, 31(28):3516–36, 2012.

[166] Sarah Donegan, Paula Williamson, Umberto D'Alessandro, Paul Garner, and Catrin Tudur Smith. Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: Individual patient data may be beneficial if only for a subset of trials. *Statistics in medicine*, 32(6):914–30, 2013.

[167] Nicky J Welton and Anthony E Ades. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 172(1):119–136, 2009.

[168] Dimitris Mavridis, Alex Sutton, Andrea Cipriani, and Georgia Salanti. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Statistics in medicine*, 32(1):51–66, 2013.

[169] Susanne Schmitz and Cathal Walsh. Incorporating data from various trial designs into a mixed treatment comparison model. *Statistics in medicine*, 32(17):2935–2949, 2013.

[170] Howard HZ Thom, Gorana Capkun, Annamaria Cerulli, Richard M Nixon, and Luke S Howard. Network meta-analysis combining individual patient and aggregate data from a mixture of study designs with an application to pulmonary arterial hypertension. *BMC Medical Research Methodology*, 15(1):34, 2015.

[171] Beth S Woods, Neil Hawkins, and David A Scott. Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: a tutorial. *BMC medical research methodology*, 10(1):54, 2010.

[172] E Moreno, F J Vázquez-Polo, and M a Negrín. Objective Bayesian meta-analysis for sparse discrete data. *Statistics in medicine*, 33(21):3676–3692, 2014.

[173] Ou Bai, Min Chen, and Xinlei Wang. Bayesian Estimation and Testing in Random Effects Meta-Analysis of Rare Binary Adverse Events. *Statistics in Biopharmaceutical Research*, 8(1):49–59, 2016.

[174] F J Vázquez, E Moreno, M A Negrín, and M Martel. Bayesian robustness in meta-analysis for studies with zero responses. *Pharmaceutical Statistics*, 15(3):230–237, 2016.

[175] Yuanyuan Tang, Qi Tang, Yao Yu, and Shihua Wen. A Bayesian Meta-analysis Method for Estimating Risk Difference of Rare Events. *Journal of Biopharmaceutical Statistics*, 28(3):550–561, 2018.

[176] D E Warn, S G Thompson, and D J Spiegelhalter. Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Statistics in medicine*, 21(11):1601–23, 2002.

[177] European Medicines Agency. Bronchitol (mannitol) - Assessment report. Technical report.

[178] Konstantinos Pateras, Stavros Nikolakopoulos, and Kit CB Roes. Prior distributions for variance parameters in a sparse-event meta-analysis of a few small trials. *Pharmaceutical Statistics*, Second Revision.

[179] European Medicines Agency. Torisel (temsirolimus) - Scientific discussion. Technical report, 2007.

[180] Heinz Schmidli, Simon Wandel, and Beat Neuenschwander. The network meta-analytic-predictive approach to non-inferiority trials. *Statistical methods in medical research*, 22(2):219–40, 2011.

[181] Adnan Y Manzur, Thierry Kuntzer, Mike Pike, and Anthony V Swan. Glucocorticoid corticosteroids for Duchenne muscular dystrophy (Review). *The Cochrane Collaboration*, (1):1–72, 2008.

[182] Usha R ani Somaraju and Marcus Merrin. Sapropterin dihydrochloride for phenylketonuria. *The Cochrane database of systematic reviews*, (3):CD008005, 2015.

[183] David R Cox. Combination of data. *Encyclopedia of statistical sciences*, 1982.

[184] John P A Ioannidis, Nikolaos a Patsopoulos, and Hannah R Rothstein. Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ*, 336(7658):1413–1415, 2008.

[185] Martyn Plummer. JAGS: A program for analysis of bayesian graphical models using gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003).*, pages 20–22, 2003.

[186] Stan Development Team. Stan: A C++ Library for Probability and Sampling, Version 2.8.0, 2015.

[187] Johanna Useem, Alana Brennan, Michael LaValley, Michelle Vickery, Omid Ameli, Nichole Reinen, and Christopher J. Gill. Systematic differences between cochrane and non-cochrane meta-analyses on the same topic: A matched pair analysis. *PLoS ONE*, 10(12):1–17, 2015.

[188] Iain Chalmers, Michael B Bracken, Ben Djulbegovic, Silvio Garattini, Jonathan Grant, A Metin Gülmezoglu, David W Howells, John P A Ioannidis, and Sandy Oliver. How to increase value and reduce waste when research priorities are set, 2014.

[189] Adriani Nikolakopoulou, Dimitris Mavridis, Matthias Egger, and Georgia Salanti. Continuously updated network meta-analysis and statistical monitoring for timely decision-making. *Statistical Methods in Medical Research*, 27(5):1312–1330, 2018.

[190] Perrine Créquit, Ludovic Trinquart, Amélie Yavchitz, and Philippe Ravaud. Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: the example of lung cancer. *BMC medicine*, 14(1):8, 2016.

[191] Per Olav Vandvik, Romina Brignardello-Petersen, and Gordon H Guyatt. Living cumulative network meta-analysis to reduce waste in research: A paradigmatic shift for systematic reviews? *BMC medicine*, 14(1):59, 2016.

[192] Pablo E. Verde and Christian Ohmann. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Research Synthesis Methods*, 6(April):45–62, 2015.

[193] Hui Yao, Ming-hui Chen, and Chunfu Qiu. Bayesian Modeling and Inference for Meta-Data with Applications in Efficacy Evaluation of an Allergic Rhinitis Drug. *Journal of Biopharmaceutical Statistics*, 21(5):992–1005, 2011.

[194] Clinical Study Data Request, url: https://www.clinicalstudydatarequest.com/, Accessed: 12/3/2020.

[195] Michelle M Mello, Jeffrey K Francer, Marc Wilenzick, Patricia Teden, Barbara E Bierer, and Mark Barnes. Preparing for Responsible Sharing of Clinical Trial Data. *The New England Journal of Medicine*, 367(27):1651–1658, 2013.

[196] David R Cox and David V Hinkley. *Theoretical statistics*. CRC Press, 1979.

[197] Tim Friede, Christian Röver, Simon Wandel, and Beat Neuenschwander. Meta-analysis of few small studies in orphan diseases. *Research Synthesis Methods*, 8(1):79–91, 2017.

[198] Joachim Hartung and Guido Knapp. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, 20(24):3875–3889, 2001.

[199] Joanna Inthout, John P A Ioannidis, George F. Borm, and Jelle J. Goeman. Small studies are more heterogeneous than large ones: A meta-meta-analysis. *Journal of Clinical Epidemiology*, 68(8):860–869, 2015.

[200] Areti Angeliki Veroniki, Dimitris Mavridis, Julian PT Higgins, and Georgia Salanti. Characteristics of a loop of evidence that affect detection and estimation of inconsistency: a simulation study. *BMC Medical Research Methodology*, 14(1):106, 2014.

[201] Stephen Senn. Trying to be precise about vagueness. *Statistics in Medicine*, 26(7):1417–1430, 2007.

[202] Nicola D. Crins, Christian Röver, Armin D. Goralczyk, and Tim Friede. Interleukin-2 receptor antagonists for pediatric liver transplant recipients: A systematic review and meta-analysis of controlled studies. *Pediatric Transplantation*, 18(8):839–850, 2014.

[203] Yan Zeng, Xin Duan, Jiajun Xu, and Xun Ni. TPO receptor agonist for chronic idiopathic thrombocytopenic purpura. In Xin Duan, editor, *Cochrane Database of Systematic Reviews*, number 7. John Wiley & Sons, Ltd, Chichester, UK, 2011.

[204] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.

[205] Dean Langan, Julian P T Higgins, and Mark Simmonds. Comparative performance of heterogeneity variance estimators in meta-analysis : a review of simulation studies. *Research Synthesis Methods*, 2015.

[206] Areti Angeliki Veroniki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian Pt Higgins, Dean Langan, and Georgia Salanti. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1):55–79, 2016.

[207] Rac Hughes, Av Swan, and Pa Van Doorn. Intravenous immunoglobulin for Guillain-Barr{é} syndrome (Review). *The cochrane Collaboration*, (9):66, 2014.

[208] Larry C Lands and Sanja Stanojevic. Oral non-steroidal anti-inflammatory drug therapy for lung disease in cystic fibrosis. Number 9. 2019.

[209] David R Cox. The Continuity Correction. *Biometrika*, 57(1):217–219, 1970.

[210] Steven E Nissen and Kathy Wolski. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *The New England journal of medicine*, 356(24):2457–2471, 2007.

[211] Michael J Sweeting, Alexander J Sutton, and Paul C Lambert. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, 23(9):1351–1375, 2004.

[212] Jonathan J Shuster, Lynn S Jones, and Daniel A Salmon. Fixed vs random effects meta-analysis in rare event studies: The Rosiglitazone link with myocardial infarction and cardiac death. *Statistics in Medicine*, 26(24):4375–4385, 2007.

[213] Gerta Rücker, Guido Schwarzer, James Carpenter, and Ingram Olkin. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine*, 28(5):721–738, 2009.

[214] Michael J Bradburn, Jonathan J Deeks, Jesse A Berlin, and Russell A Localio. Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, 26(1):53–77, 2007.

[215] Evangelos Kontopantelis, David A Springate, and David Reeves. A Re-Analysis of the Cochrane Library Data: The Dangers of Unobserved Heterogeneity in Meta-Analyses. *PLoS ONE*, 8(7):1–14, 2013.

[216] Robert C Paule and John Mandel. Consensus Values And Weighting Factors. *Journal of Research of the National Bureau of Standards*, 87(5):377–385, 1982.

[217] Kurex Sidik and Jeffrey N Jonkman. Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 54(2):367–384, 2005.

[218] J Hartung and K H Makambi. Reducing the number of unjustified significant results in meta-analysis. *Communications in Statistics-Simulation and Computation*, 32(4):1179–1190, 2003.

[219] John E Hunter and Frank L Schmidt. *Methods of meta-analysis: Correcting error and bias in research findings*. Sage, 2004.

[220] Larry V Hedges and Ingram Olkin. *Statistical methods for meta-analysis*. Academic press, ., 2014.

[221] Kurex Sidik and Jeffrey N. Jonkman. A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, 26(9):1964–1981, 2007.

[222] W. Viechtbauer. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293, 2005.

[223] Dean Langan, Julian P T Higgins, and Mark Simmonds. An empirical comparison of heterogeneity variance estimators in 12 894 meta-analyses. *Research synthesis methods*, 6(2):195–205, 2015.

[224] Orphanet. "Website", Accessed: 2020-03-12.

[225] Annamaria Guolo and Cristiano Varin. Random-effects meta-analysis: the number of studies matters. *Statistical Methods in Medical Research*, 26(3):1500–1518, 2017.

[226] Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G Ibrahim, Nelson Kinnersley, Stacy Lindborg, Sandrine Micallef, Satrajit Roychoudhury, and Laura Thompson. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics*, 13(1):41–54, 2014.

[227] Fiona C Warren, Keith R Abrams, and Alex J Sutton. Hierarchical network meta-analysis models to address sparsity of events and differing treatment classifications with regard to adverse outcomes. *Statistics in medicine*, 33(14):2449–2466, 2014.

[228] Ivo N van Schaik, Leonard H van den Berg, Rob de Haan, and Marinus Vermeulen. Intravenous immunoglobulin for multifocal motor neuropathy (Review). *Cochrane Database of Systematic Reviews*, (2), 2005.

[229] Burak Kürsad Günhan, Christian Röver, and Tim Friede. *Research Synthesis Methods*.

[230] William J Browne and David Draper. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514, 2006.

[231] Małgorzata Roos and Leonhard Held. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6(2):259–278, 2011.

[232] George E P Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

[233] Elizabeth Stojanovski and KerrieL Mengersen. Bayesian Methods in Meta-Analysis. In *Encyclopedia of Biopharmaceutical Statistics, Third Edition*, pages 116–121. CRC Press, 2012.

[234] Dan Jackson, Martin Law, Theo Stijnen, Wolfgang Viechtbauer, and Ian R White. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in medicine*, 37(7):1059–1085, 2018.

[235] Nicholas G Polson and James G Scott. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

[236] William DuMouchel and Sharon-Lise Normand. Computer-modeling and graphical strategies for meta-analysis. *Meta-analysis in medicine and health policy*, pages 119–164, 2000.

[237] Mj Daniels. A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, 27(3):567–578, 1999.

[238] Orphanet: About Rare Diseases. http://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=EN, Accessed: 2020-03-12.

[239] European Medicines Agency. Kiovig (Motor neuropathy) - Assessment report. Technical report, 2011.

[240] M Plummer and A Stukalov. Package 'rjags', 2015.

[241] Konstantinos Pateras, Stavros Nikolakopoulos, and Kit CB Roes. Data-generating models of dichotomous outcomes: Heterogeneity in simulation studies for a random-effects meta-analysis. *Statistics in medicine*, 37(7):1115–1124, 2018.

[242] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[243] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6(1):7–11, 2006.

[244] Lie Li and Xinlei Wang. Meta-Analysis of Rare Binary Events in Treatment Groups with Unequal Variability. *Statistical Methods in Medical Research*, 28(1):263–274, 2019.

[245] Frédéric Cérou, Pierre Del Moral, Teddy Furon, and Arnaud Guyader. Sequential Monte Carlo for rare event estimation. *Statistics and Computing*, 22(3):795–808, 2012.

[246] Zdravko I Botev and Dirk P Kroese. Efficient Monte Carlo simulation via the generalized splitting method. *Statistics and Computing*, 22(1):1–16, 2012.

[247] European Medicines Agency. Thalidomide celgene (previously thalidomide pharmion, thalidomide) - assessment report. pages 1–91, 2008.

[248] European Medicines Agency. Galafold (migalastat) - assesment report. pages 1–91, 2012.

[249] Raphael Schiffmann, Markus Ries, Derek Blankenship, Kathy Nicholls, Atul Mehta, Joe TR Clarke, Robert D Steiner, Michael Beck, Bruce A Barshop, William Rhead, et al. Changes in plasma and urine globotriaosylceramide levels do not predict fabry disease progression over 1 year of agalsidase alfa. *Genetics in Medicine*, 15(12):983, 2013.

[250] B Engel and P Walstra. Increasing precision or reducing expense in regression experiments by using information from a concomitant variable. *Biometrics*, pages 13–20, 1991.

[251] D Conniffe and MA Moran. Double sampling with regression in comparative studies of carcass composition. *Biometrics*, pages 1011–1023, 1972.

[252] Cornelia Ursula Kunz, Tim Friede, Nicholas Parsons, Susan Todd, and Nigel Stallard. A comparison of methods for treatment selection in seamless phase ii/iii clinical trials incorporating information on short-term endpoints. *Journal of biopharmaceutical statistics*, 25(1):170–189, 2015.

[253] Nigel Stallard. A confirmatory seamless phase ii/iii clinical trial design incorporating short-term endpoint information. *Statistics in medicine*, 29(9):959–971, 2010.

[254] Lisa V Hampson and Christopher Jennison. Optimizing the data combination rule for seamless phase ii/iii clinical trials. *Statistics in medicine*, 34(1):39–58, 2015.

[255] David Manner, John W Seaman Jr, and Dean M Young. Bayesian methods for regression using surrogate variables. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 46(6):750–759, 2004.

[256] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC, 2014.

[257] John Barnard, Robert McCulloch, and Xiao-Li Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311, 2000.

[258] Joseph G Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. On optimality properties of the power prior. *Journal of the American Statistical Association*, 98(461):204–213, 2003.

[259] Stavros Nikolakopoulos, Ingeborg van der Tweel, and Kit CB Roes. Dynamic borrowing through empirical power priors that control type i error. *Biometrics*, 74(3):874–880, 2018.

[260] Isaac Gravestock, Leonhard Held, and COMBACTE-Net consortium. Adaptive power priors with empirical bayes for clinical trials. *Pharmaceutical statistics*, 16(5):349–360, 2017.

[261] Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

[262] Joseph G Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in medicine*, 34(28):3724–3749, 2015.

[263] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[264] Donald R Barr and E Todd Sherrill. Mean and variance of truncated normal distributions. *The American Statistician*, 53(4):357–361, 1999.

[265] Uwe Malzahn, Dankmar Böhning, and Heinz Holling. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika*, 87(3):619–632, 2000.

[266] Norman Lloyd Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. Continuous univariate distributions. 1994.

[267] Loukia M Spineli, Nikolaos Pandis, and Georgia Salanti. Reporting and handling missing outcome data in mental health: a systematic review of cochrane systematic reviews and meta-analyses. *Research synthesis methods*, 6(2):175–187, 2015.

[268] Lara A Kahale, Batoul Diab, Romina Brignardello-Petersen, Arnav Agarwal, Reem A Mustafa, Joey Kwong, Ignacio Neumann, Ling Li, Luciane Cruz Lopes, Matthias Briel, et al. Systematic reviews do not adequately report or address missing outcome data in their analyses: a methodological survey. *Journal of clinical epidemiology*, 99:14–23, 2018.

[269] Elie A Akl, Alonso Carrasco-Labra, Romina Brignardello-Petersen, Ignacio Neumann, Bradley C Johnston, Xin Sun, Matthias Briel, Jason W Busse, Shanil Ebrahim, Carlos E Granados, et al. Reporting, handling and assessing the risk of bias associated with

missing participant data in systematic reviews: a methodological survey. *BMJ open*, 5(9):e009368, 2015.

[270] Loukia M Spineli. Missing binary data extraction challenges from cochrane reviews in mental health and campbell reviews with implications for empirical research. *Research synthesis methods*, 8(4):514–525, 2017.

[271] Elie A Akl, Bradley C Johnston, Pablo Alonso-Coello, Ignacio Neumann, Shanil Ebrahim, Matthias Briel, Deborah J Cook, and Gordon H Guyatt. Addressing dichotomous data for participants excluded from trial analysis: a guide for systematic reviewers. *PloS one*, 8(2):e57132, 2013.

[272] Julian PT Higgins, Ian R White, and Angela M Wood. Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical Trials*, 5(3):225–239, 2008.

[273] NL Turner, Sofia Dias, Anthony E Ades, and Nicky J Welton. A bayesian framework to account for uncertainty due to missing binary outcome data in pairwise meta-analysis. *Statistics in medicine*, 34(12):2062–2080, 2015.

[274] Loukia M Spineli. Modeling missing binary outcome data while preserving transitivity assumption yielded more credible network meta-analysis results. *Journal of clinical epidemiology*, 105:19–26, 2019.

[275] Henry OD Ejere, Ellen Schwartz, Richard Wormald, and Jennifer R Evans. Ivermectin for onchocercal eye disease (river blindness). *Cochrane Database of Systematic Reviews*, (8), 2012.

[276] Evan Mayo-Wilson, Susan Hutfless, Tianjing Li, Gillian Gresham, Nicole Fusco, Jeffrey Ehmsen, James Heyward, Swaroop Vedula, Diana Lock, Jennifer Haythornthwaite, et al. Integrating multiple data sources (muds) for meta-analysis to improve patient-centered outcomes research: a protocol for a systematic review. *Systematic reviews*, 4(1):143, 2015.

[277] Gianni Virgili, Manuele Michelessi, Maurizio B Parodi, Daniela Bacherini, and Jennifer R Evans. Laser treatment of drusen to prevent progression to advanced age-related macular degeneration. *Cochrane Database of Systematic Reviews*, (10), 2015.

[278] Jennifer Watt, Zahra Goodarzi, Andrea C Tricco, Areti-Angeliki Veroniki, and Sharon E Straus. Comparative safety and efficacy of pharmacological and non-pharmacological interventions for the behavioral and psychological symptoms of dementia: protocol for a systematic review and network meta-analysis. *Systematic reviews*, 6(1):182, 2017.

[279] Areti Angeliki Veroniki, Sharon E Straus, Huda M Ashoor, Jemila S Hamid, Brenda R Hemmelgarn, Jayna Holroyd-Leduc, Sumit R Majumdar, Glenn McAuley, and Andrea C Tricco. Comparative safety and effectiveness of cognitive enhancers for alzheimer's dementia: protocol for a systematic review and individual patient data network meta-analysis. *BMJ open*, 6(1), 2016.

[280] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

[281] Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

[282] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[283] Bart Michiels, Geert Molenberghs, and Stuart R Lipsitz. Selection models and pattern-mixture models for incomplete data with covariates. *Biometrics*, 55(3):978–983, 1999.

[284] Ian R White, Nicky J Welton, Angela M Wood, AE Ades, and Julian PT Higgins. Allowing for uncertainty due to missing data in meta-analysis—part 2: hierarchical models. *Statistics in medicine*, 27(5):728–745, 2008.

[285] Loukia M Spineli, Julian PT Higgins, Andrea Cipriani, Stefan Leucht, and Georgia Salanti. Evaluating the impact of imputations for missing participant outcome data in a network meta-analysis. *Clinical Trials*, 10(3):378–388, 2013.

[286] Orestis Efthimiou, Thomas PA Debray, Gert van Valkenhoef, Sven Trelle, Klea Panayidou, Karel GM Moons, Johannes B Reitsma, Aijing Shang, Georgia Salanti, and GetReal Methods Review Group. Getreal in network meta-analysis: a review of the methodology. *Research synthesis methods*, 7(3):236–263, 2016.

[287] Maria Petropoulou, Adriani Nikolakopoulou, Areti-Angeliki Veroniki, Patricia Rios, Afshin Vafaei, Wasifa Zarin, Myrsini Giannatsi, Shannon Sullivan, Andrea C Tricco, Anna Chaimani, et al. Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. *Journal of clinical epidemiology*, 82:20–28, 2017.

[288] Peter Diggle and Michael G Kenward. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):49–73, 1994.

[289] Ian R White, James Carpenter, and Nicholas J Horton. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clinical trials*, 9(4):396–407, 2012.

[290] Loukia M Spineli. An empirical comparison of bayesian modelling strategies for missing binary outcome data in network meta-analysis. *BMC medical research methodology*, 19(1):86, 2019.

[291] Sofia Dias, Alex J Sutton, AE Ades, and Nicky J Welton. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making*, 33(5):607–617, 2013.

[292] Angelo J Franchini, Sofia Dias, Anthony E Ades, Jeroen P Jansen, and Nicky J Welton. Accounting for correlation in network meta-analysis with multi-arm trials. *Research synthesis methods*, 3(2):142–160, 2012.

[293] Guobing Lu and AE Ades. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association*, 101(474):447–459, 2006.

[294] Georgia Salanti, AE Ades, and John PA Ioannidis. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of clinical epidemiology*, 64(2):163–171, 2011.

[295] Sofia Dias, Nicky J Welton, DM Caldwell, and Anthony E Ades. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in medicine*, 29(7-8):932–944, 2010.

[296] Ian R White, Julian PT Higgins, and Angela M Wood. Allowing for uncertainty due to missing data in meta-analysis—part 1: two-stage methods. *Statistics in medicine*, 27(5):711–727, 2008.

[297] Julia M Bottomley, Rod S Taylor, and Jacob Ryttov. The effectiveness of two-compound formulation calcipotriol and betamethasone dipropionate gel in the treatment of moderately severe scalp psoriasis: a systematic review of direct and indirect evidence. *Current medical research and opinion*, 27(1):251–268, 2011.

[298] Jinling Liu, Jiangchuan Dong, Lei Wang, Ying Su, Peng Yan, and Shenggang Sun. Comparative efficacy and acceptability of antidepressants in parkinson's disease: a network meta-analysis. *PloS one*, 8(10), 2013.

[299] Ariel Dogliotti, Ernesto Paolasso, and Robert P Giugliano. Current and new oral antithrombotics in non-valvular atrial fibrillation: a network meta-analysis of 79 808 patients. *Heart*, 100(5):396–405, 2014.

[300] David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337, 2000.

[301] Georg Kemmler, Martina Hummer, Christian Widschwendter, and W Wolfgang Fleischhacker. Dropout rates in placebo-controlled and active-control clinical trials of antipsychotic drugs: a meta-analysis. *Archives of General Psychiatry*, 62(12):1305–1312, 2005.

[302] Loukia M Spineli, Stefan Leucht, Andrea Cipriani, Julian PT Higgins, and Georgia Salanti. The impact of trial characteristics on premature discontinuation of antipsychotics in schizophrenia. *European Neuropsychopharmacology*, 23(9):1010–1016, 2013.

[303] Stef Van Buuren, Hendriek C Boshuizen, and Dick L Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18(6):681–694, 1999.

[304] James R Carpenter and Michael G Kenward. Missing data in randomised controlled trials: a practical guide, 2007.

[305] Beat Neuenschwander, Gorana Capkun-Niggli, Michael Branson, and David J Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical trials (London, England)*, 7(1):5–18, 2010.

[306] Brian P Hobbs, Daniel J Sargent, and Bradley P Carlin. Commensurate Priors for Incorporating Historical Information in Clinical Trials Using General and Generalized Linear Models. *Bayesian Analysis*, (3):639–674, 2012.

[307] Svenja E Seide, Christian Röver, and Tim Friede. Likelihood-based random-effects meta-analysis with few studies: empirical and simulation studies. *BMC medical research methodology*, 19(1):16, 2019.

[308] Tim Mathes and Oliver Kuss. A comparison of methods for meta-analysis of a small number of studies with binary outcomes. *Research synthesis methods*, 9(3):366–381, 2018.

[309] Christopher Partlett and Richard D Riley. Random effects meta-analysis : Coverage performance of 95% confidence and prediction intervals following REML estimation. 36(2):301–317, 2017.

[310] Marie Beisemann, Philipp Doebler, and Heinz Holling. Comparison of random-effects meta-analysis models for the relative risk in the case of rare events: A simulation study. *Biometrical Journal*, 2020.

[311] Svenja E Seide, Katrin Jensen, and Meinhard Kieser. Simulation and data-generation for random-effects network meta-analysis of binary outcome. *Statistics in medicine*, 38(17):3288–3303, 2019.

[312] Loukia M Spineli, Chrysostomos Kalyvas, and Konstantinos Pateras. Participants' outcomes gone missing within a network of interventions: Bayesian modeling strategies. *Statistics in medicine*, 38(20):3861–3879, 2019.

[313] Loukia M Spineli and Chrysostomos Kalyvas. Comparison of exclusion, imputation and modelling of missing binary outcome data in frequentist network meta-analysis. *BMC Medical Research Methodology*, 20(1):1–15, 2020.

[314] Chrysostomos Kalyvas, Loukia M Spineli, and Konstantinos Pateras. Caution required in the analysis of poorly connected networks of intervention. *International Society of Clinical Biostatistics*, July 2019.

[315] "European Parliament and European Council". Regulation (EC) No 141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products . pages 1–5, 2010.

[316] Segolene Ayme. Discussion on evidence synthesis; workshop on small population studies, European Medicines Agency, March 2017.

Summary

English summary

"Patients in rare diseases should be entitled to the same quality of treatment as other patients" [315]. The recent implementation of international and national strategy plans for rare diseases has led to an incredible increase of data quantity. As is generally known, such data become available through various sources, for instance; claim databases, e-health records, national registries and experimental studies. Information provided from either of these sources can be utilized during the evaluation of a (new) treatment.

Depending on the stakeholder, we often need to evaluate multiple questions; such as, (i) do the treatment's risks outweigh the added benefit?, (ii) does the treatment's added benefit outweigh the costs?, (iii) does the treatment help each and every individual patient or (iv) does the treatment actually work? The latter question is often evaluated by a randomized controlled trial. According to Orphanet, more than 1829 trials among 29 countries focus on more than 800 rare conditions [316]. This number of historical and ongoing trials explains the urgent need for evaluating and developing tailor-made statistical methods for small populations and so far, three recent large European projects have responded to materializing this need [6, 7, 8].

Statisticians, methodologists, clinicians and patients, all interested parties recognize the specific issue of the increased presence of heterogeneity among the patients that suffer from a specific rare condition. Heterogeneity introduces further complexity during the evaluation of a (new) treatment within a single randomized clinical trial, where the limited availability of sample sizes restricts practitioners from evaluating subgroups of patients within each trial. Although counter-intuitive, as sample sizes per trial become smaller, conducting more than one trial might provide a clearer overview of the treatment efficacy [89].

Orphan drugs for rare diseases are often investigated through multinational randomized controlled trials. Investigations in such a global setting may lead to increased inconsistency, given that the clinical expertise, the standards of care and the used facilities vary e.g. between each country. Non-homogeneous data makes single trials less reliable, which means that the need to explore synthesis methods of a few small studies through meta-analysis is even more necessary. This thesis focuses on the latter.

More specifically, Chapters 2a and 2b report scoping reviews that identified (i) recent

developments in frequentist and Bayesian evidence synthesis of small populations and (ii) opportunities during an orphan drug authorization, where such methods could offer attractive alternatives to the current practice. One of the main findings of these reviews was that meta-analytical methods tailored for small population meta-analyses have not been thoroughly researched as of then. The lack of sufficient existing methods, combined with their improper evaluation is considered the underlying rationale for this thesis which hopefully provides grounds for further research development in the area.

Chapter 3 focuses on the binomial data generating mechanisms that are utilized to evaluate the performance of meta-analytical methods. We have showed that often no rationale exists as regards to the choice of such models among individual simulation studies. When alternative data-generating mechanisms are applied, we have observed heavy discrepancies among the evaluated statistical methods performance, especially in situations of a synthesis of a few small trials.

Chapter 4 deals with the problematic estimation of heterogeneity when synthesizing a few small trials with reported zero events. Such a setting offered the opportunity to compare the behaviour of heterogeneity estimation techniques under the exact number of observed zero events. When accounting for the exact number of observed zero events in small population settings, we found no obvious discrepancies among the applied techniques.

Chapter 5 builds on the problematic nature of frequentist random-effects methods, which were previously explored in Chapter 4, now assuming a Bayesian binomial-normal random-effects model. We applied alternative prior distributions on the heterogeneity and we observed deviations both on the point and the interval estimation of the overall effect. If possible given the clinical question at hand, our recommendation is that priors on the heterogeneity parameter, except for clinically relevant, they should neither be non-informative nor very informative. Given the above recommendation, inferences of sparse-events meta-analyses should provide sensible but not prior-driven inferences.

Chapter 6 focuses on the situation where only one phase II and one phase III trial are available to assess the efficacy of a (new) intervention. Due to limited time and overall resources, phase II clinical trials may only observe short-term outcomes, while the Phase III trial often contains

results for both the short and long-term outcomes, thus, quantification of their relationship and borrowing of strength could take place when appropriate based on statistical approaches suggested in this Chapter.

Chapter 7 explores the behaviour of different proposed and/or existing Bayesian modelling options for handling binomial missing-outcome data in a network meta-analysis. The overall suggestion is that, prior to conducting a network meta-analysis, researchers should set a sensible prior assumption on the missing outcome mechanism. We expect the prior specification to be a far more crucial matter within a sparsely connected meta-analysis network of a few small trials.

Περίληψη στα Ελληνικά

Περίληψη στα Ελληνικά

¨Οι ασθενείς με σπάνια νοσήματα δικαιούνται την ίδια θεραπευτική ποιότητα όπως και οι υπόλοιποι ασθενείς.' [315]. Η πρόσφατη υλοποίηση διεθνών και εθνικών στρατηγικών σχεδίων για τα σπάνια νοσήματα έχει φέρει μία μεγάλη αύξηση στο μέγεθος των διαθέσιμων δεδομένων. Όπως είναι γνωστό, τέτοια δεδομένα γίνονται διαθέσιμα μέσω διαφόρων πηγών, για παράδειγμα· βάσεις δεδομένων, ηλεκτρονικά μητρώα υγείας, εθνικά μητρώα αλλά και κλινικές πειραματικές μελέτες. Πληροφορία που προέρχεται από οποιαδήποτε από τις παραπάνω πηγές μπορεί να χρησιμοποιηθεί κατά την αξιολόγηση μίας (νέας) θεραπευτικής αγωγής.

Ανάλογα τα ενδιαφερόμενα μέρη, συνήθως αξιολογούμε πολλαπλά ερωτήματα όπως, (i) οι κίνδυνοι λήψης της φαρμακευτικής αγωγής υπερισχύουν του πρόσθετου κέρδους;, (ii) το κέρδος της λήψης θεραπείας υπερισχύει του οικονομικού βάρους;, (iii) βοηθάει η θεραπεία όλους τους ασθενείς; ή/και (iv) λειτουργεί όντως η νέα θεραπεία; Το τελευταίο ερώτημα συχνά αξιολογείται με τη χρήση μίας τυχαιοποιημένης κλινικής δοκιμής. Σύμφωνα με τη βάση *Orphanet*, τρέχουν περισσότερες από 1829 κλινικές δοκιμές μεταξύ 29 χωρών για περισσότερα από 800 φάρμακα [316]. Αυτός ο αριθμός ιστορικών και τρεχόντων κλινικών δοκιμών εξηγεί την άμεση ανάγκη να αξιολογηθούν και να υλοποιηθούν νέες στοχευμένες στατιστικές μέθοδοι για μικρούς πληθυσμούς. Τουλάχιστον τρία μεγάλα ευρωπαϊκά προγράμματα υλοποιήθηκαν για να καλύψουν αυτή την ανάγκη [6, 7, 8].

Στατιστικοί, επιδημιολόγοι, ιατροί και ασθενείς, όλα τα ενδιαφερόμενα μέρη, αναγνωρίζουν το θέμα της αυξημένης ετερογένειας ανάμεσα σε ασθενείς που πάσχουν από κάποιο σπάνιο νόσημα. Η ετερογένεια αυτή φέρνει επιπλέον πολυπλοκότητα κατά την αξιολόγηση μίας (νέας) θεραπείας μέσα σε μία μοναδική τυχαιοποιημένη κλινική δοκιμή, όπου το μέγεθος του δείγματος περιορίζει τους ερευνητές να πραγματοποιήσουν ανάλυση σε υποομάδες ασθενών.

Τα ορφανά φάρμακα στα σπάνια νοσήματα συνήθως ερευνούνται μέσω μικρών κλινικών δοκιμών ή μέσω μεγαλύτερων πολυεθνικών τυχαιοποιημένων κλινικών δοκιμών. Η έρευνα σε τέτοιο παγκόσμιο επίπεδο μπορεί να επιφέρει αύξηση της ασυνέπειας και της ετερογένειας εφόσον, π.χ. η κλινική γνώση, η καθιερωμένη φροντίδα αλλά και οι υπάρχουσες εγκαταστάσεις διαφέρουν από χώρα σε χώρα. Τέτοιου είδους μη ομογενοποιημένα δεδομένα μειώνουν την εμπιστοσύνη των ερευνητών σε μεμονωμένες (μικρές) κλινικές δοκιμές, κάτι που σημαίνει ότι η σύνθεση τέτοιων μελετών μέσω μίας μέτα-ανάλυσης είναι αναγκαία. Η διατριβή αυτή συγκεντρώνεται στο τελευταίο σημείο εκ των παραπάνω.

Συγκεκριμένα, τα κεφάλαια 2α και 2β παρουσιάζουν διερευνητικές ανασκοπήσεις οι οποίες εντόπισαν (*i*) πρόσφατες εξελίξεις σε κλασσικές και Μπεϋζιανές στατιστικές μεθόδους σύνθεσης δεδομένων σε μικρούς πληθυσμούς και (*ii*) ευκαιρίες που παρουσιάζονται εντός μίας αξιολόγησης ορφανού φαρμάκου, στις οποίες τέτοιες μέθοδοι μπορούν να προσφέρουν εναλλακτικές σε σύγκριση με τη πάγιες τακτικές. Ένα από τα κύρια ευρήματα των ανασκοπήσεων είναι ότι μετα-αναλυτικές μέθοδοι φτιαγμένοι για μικρούς πληθυσμούς δεν είχαν αρχίσει να ερευνούνται με συστηματικότητα μέχρι τότε. Η έλλειψη επαρκών τεχνικών, όπως επίσης και ο συχνά ακατάλληλος τρόπος αξιολόγησής τους είναι τα θέματα στα οποία βασίστηκε αυτή η διατριβή αυτή.

Το κεφάλαιο 3 επικεντρώνεται σε διωνυμικούς μηχανισμούς παραγωγής δεδομένων οι οποίοι χρησιμοποιούνται για να αξιολογήσουν την απόδοση μετα-αναλυτικών μεθόδων. Παρουσιάζεται ότι συχνά δεν υπάρχει επαρκής δικαιολόγηση για την επιλογή τέτοιων μηχανισμών μεταξύ των μελετών προσομοίωσης. Όταν διαφορετικοί μηχανισμοί εφαρμόζονται παρατηρούνται ουσιαστικές διαφοροποιήσεις στην απόδοση των στατιστικών μεθόδων που αξιολογούνται, αυτές οι διαφοροποιήσεις ήταν έντονες σε περιπτώσεις σύνθεσης λίγων μικρών κλινικών δοκιμών.

Το κεφάλαιο 4 αντιμετωπίζει το θέμα της προβληματικής εκτίμησης της ετερογένειας όταν συνθέτονται δυωνυμικά καταληκτικά σημεία λίγων μικρών κλινικών δοκιμών με μηδενικά παρατηρηθέντα γεγονότα (zero events). Οι συνθήκες μικρών πληθυσμών προσέφεραν την ευκαιρία να συγκρίνουμε τη συμπεριφορά των τεχνικών εκτίμησης ετερογένειας σε σχέση με τον ακριβή αριθμό παρατηρηθέντων μηδενικών γεγονότων (zero events). Μεταξύ διαφορετικών συχνοτήτων ύπαρξης μηδενικών, σε μία τέτοια μέτα-ανάλυση μικρών πληθυσμών, δεν εντοπίστηκαν εμφανείς αποκλίσεις μεταξύ των εφαρμοζόμενων τεχνικών.

Το Κεφάλαιο 5 χτίζει πάνω στην προβληματική φύση των κλασσικών μεθόδων τυχαίων επιδράσεων, οι οποίες είχαν προηγουμένως διερευνηθεί στο Κεφάλαιο 4, τώρα υποθέτοντας ένα Μπεϋζιανό διωνυμικό-κανονικό μοντέλο τυχαίων επιδράσεων. Εφαρμόστηκαν εναλλακτικές εκ των προτέρων κατανομές πάνω στην παράμετρος της ετερογένειας και παρατηρήθηκαν αποκλίσεις τόσο στο σημειακό εκ των υστέρων διάμεσο της παραμέτρου όσο και στο διάστημα αξιοπιστίας του. Σε αυτό το κεφάλαιο προτάθηκε, δεδομένης της κλινικής ερώτησης, να γίνεται χρήση εκ των προτέρων κατανομών ετερογένειας οι οποίες δεν είναι πολύ αλλά ούτε και λίγο

πληροφοριακές. Με τη χρήση των παραπάνω προτάσεων τα αποτελέσματα μίας Μπεϋζιανής μέτα-ανάλυσης θα παρέχουν λογικά αποτελέσματα, τα οποία είναι ανεξάρτητα από τις υποθέσεις στις εκ των προτέρων κατανομές ετερογένειας.

Το Κεφάλαιο 6 επικεντρώνεται στην περίπτωση όπου μόνο μία δοκιμή Φάση ΙΙ και μία δοκιμή Φάση ΙΙΙ είναι διαθέσιμες για την αξιολόγηση της αποτελεσματικότητας μιας (νέας) παρέμβασης. Λόγω του περιορισμένου χρόνου και των συνολικών πόρων, οι κλινικές δοκιμές Φάσης ΙΙ μπορούν να προσφέρουν μόνο βραχυπρόθεσμα καταληκτικά σημεία, ενώ η δοκιμή Φάσης ΙΙΙ συχνά περιέχει αποτελέσματα τόσο για τα βραχυπρόθεσμα όσο και για τα μακροπρόθεσμα καταληκτικά σημεία. Ο ποσοτικός προσδιορισμός της σχέσης των σημείων αυτών και ο δανεισμός δύναμης από τη δοκιμή Φάσης ΙΙ κατά την αξιολόγηση των ορφανών φαρμακευτικών αγωγών (orphan drugs) θα μπορούσαν να πραγματοποιηθούν όποτε ενδείκνυται βάσει των στατιστικών προσεγγίσεων που προτείνονται στο παρόν κεφάλαιο.

Το Κεφάλαιο 7 διερευνά τη συμπεριφορά προτεινόμενων ή/και υπαρχουσών Μπεϋζιανων μοντέλων για το χειρισμό διωνυμικών δεδομένων με ελλιπείς παρατηρήσεις (missing outcome data) σε μία μετα-ανάλυση δικτύου (network meta-analysis). Στο κεφάλαιο αυτό προτείνεται ότι, πριν από τη διεξαγωγή μετα-ανάλυσης δικτύου, οι ερευνητές θα πρέπει να θέσουν μια λογική εκ των προτέρων κατανομή σχετικά με τον μηχανισμό εμφάνισης ελλιπών παρατηρήσεων. Αναμένεται ότι η τοποθέτηση εκ των προτέρων κατανομής στο μηχανισμό εμφάνισης ελλιπών παρατηρήσεων θα είναι πολύ πιο καίριο ζήτημα σε ένα αραιά συνδεδεμένο δίκτυο μετα-ανάλυσης μερικών μικρών δοκιμών.

Nederlandse samenvatting

”Patiënten met zeldzame ziekten moeten recht hebben op dezelfde kwaliteit van behandeling als andere patiënten” [315]. De recente implementatie van internationale en nationale strategieplannen voor zeldzame ziekten hebben geleid tot een ongelooflijke toename van de hoeveelheid gegevens. Zoals bekend komen dergelijke gegevens via verschillende bronnen beschikbaar: bijvoorbeeld via claim-databases, e-health-records, nationale registers en experimentele studies –(klinische proeven)-. Informatie uit deze bronnen kan worden gebruikt tijdens de evaluatie van een (nieuwe) behandeling.

Afhankelijk van de belanghebbende moeten we vaak meerdere vragen evalueren, bijvoorbeeld: a) wegen de risico's van de behandeling zwaarder dan het toegevoegde voordeel?, (b) weegt het toegevoegde voordeel van de behandeling zwaarder dan de kosten?, (iii) helpt de behandeling elke individuele patiënt of (c) werkt de behandeling ook daadwerkelijk? De laatste vraag wordt vaak geëvalueerd door een gerandomiseerde gecontroleerde proef. Volgens Orphanet richten meer dan 1829 proefnemingen onder 29 landen zich op meer dan 800 zeldzame aandoeningen [316]. Het aantal historische en lopende proeven verklaart de noodzaak om op maat gemaakte statistische methoden voor kleine populaties te evalueren en te ontwikkelen. Tot nu toe hebben minstens drie grote Europese projecten gereageerd op het materialiseren van deze behoefte [6, 7, 8].

Statistici, methodologen, artsen en patiënten, alle belanghebbenden, erkennen het specifieke probleem van de toegenomen aanwezigheid van heterogeniteit bij patiënten die lijden aan een zeldzame aandoening. Heterogeniteit introduceert verdere complexiteit tijdens de evaluatie van een (nieuwe) behandeling binnen één willekeurig verdeelde klinische proef, waarbij de beperkte beschikbaarheid van steekproefomvang beoefenaars erin beperkt subgroepen patiënten binnen elke studie te evalueren.

Weesgeneesmiddelen ('Orphan Drugs') voor zeldzame ziekten worden vaak onderzocht door middel van kleine klinische proeven en/of door multinationale gerandomiseerde gecontroleerde proeven. Onderzoek in een mondiale setting kan leiden tot een grotere inconsistentie, gezien het feit dat de klinische expertise, de zorgstandaarden en de gebruikte faciliteiten b.v. per land verschillen. Niet-homogene data maakt afzonderlijke kleine proeven minder betrouwbaar, wat betekent dat de noodzaak om synthesemethoden van een paar kleine proeven te onderzoeken door middel van meta-analyse nog noodzakelijker is. Dit

proefschrift richt zich op het laatste.

Meer specifiek rapporteren hoofdstukken 2a en 2b een aantal onderzoeken die (i) recente ontwikkelingen in de frequentist en Bayesiaanse bewijssynthese van kleine populaties en (II) mogelijkheden tijdens de ontwikkeling van weesgeneesmiddelen identificeerden, waar dergelijke methoden aantrekkelijke alternatieven voor de huidige praktijk zouden kunnen bieden. Een van de belangrijkste bevindingen van deze reviews was dat meta-analytische methoden die zijn toegesneden op meta-analyses van kleine populaties, vanaf dat moment niet grondig zijn onderzocht. Het ontbreken van voldoende bestaande methoden, gecombineerd met hun ondeugdelijke evaluatie, wordt beschouwd als de onderliggende reden voor dit proefschrift.

Hoofdstuk 3 richt zich op de binomiale data genererende mechanismen die worden gebruikt om de prestaties van meta-analytische methoden te evalueren. Hoofdstuk 3 laat zien dat er vaak geen reden bestaat voor de keuze van dergelijke modellen in individuele simulatiestudies. Wanneer alternatieve gegevensgenererende mechanismen worden toegepast, worden zware verschillen tussen de geëvalueerde statistische methodeprestaties waargenomen, vooral in de synthese van een paar kleine proeven.

Hoofdstuk 4 behandelt de problematische inschatting van heterogeniteit bij het synthetiseren van binomaire resultaten van een paar kleine proeven met gerapporteerde nul-voorvallen. Zo'n setting bood de mogelijkheid om het gedrag van heterogeniteitsschattingen te vergelijken onder een exact aantal waargenomen nul-voorvallen. Bij het berekenen van het exacte aantal waargenomen nul-voorvallen in kleine populaties werden geen duidelijke verschillen gevonden tussen het geschatte behandelingseffect bij het toepassen van verschillende heterogeniteitsschattingen.

Hoofdstuk 5 bouwt voort op de problematische aard van het frequentist random-effects model voor een schaars-event meta-analyse, die eerder in hoofdstuk 4 werd onderzocht. Hoofdstuk 5 gaat nu uit van een Bayesiaans binomiaal-normaal willekeurig-effect model. In dit hoofdstuk worden alternatieve eerdere distributies toegepast op de parameter heterogeniteit en worden afwijkingen zowel op het punt als op de intervalschatting van het algehele behandelingseffect waargenomen. Indien mogelijk is de aanbeveling, gegeven de klinische vraag die hier

ter discussie staat, dat eerdere onderzoeken op de heterogene parameter, behalve klinisch relevant, niet informatief of zeer informatief mogen zijn. Gezien de bovenstaande aanbeveling kunnen de conclusies van de meta-analyses van spaargebeurtenissen verstandige maar niet van tevoren gedreven conclusies opleveren.

Hoofdstuk 6 richt zich op de situatie waarin slechts één fase II- en één fase III-proef beschikbaar is om de werkzaamheid van een (nieuwe) interventie te beoordelen. Vanwege beperkte tijd en algemene middelen kunnen klinische proeven in fase II alleen resultaten op de korte termijn waarnemen, terwijl de fase III-studie vaak resultaten bevat voor zowel de resultaten op de korte als de lange termijn, wanneer een passende kwantificering van de relatie tussen de resultaten en het lenen van kracht zou kunnen plaatsvinden op basis van de statistische benaderingen die daarin worden voorgesteld.

Hoofdstuk 7 onderzoekt het gedrag van verschillende voorgestelde en/of bestaande Bayesiaanse modelleringsopties voor het verwerken van binomiale ontbrekende-resultaatgegevens in een netwerk meta-analyse. De algemene suggestie is dat, voordat een netwerk meta-analyse wordt uitgevoerd; onderzoekers moeten een verstandige vooronderstelling geven over het ontbrekende resultaatmechanisme. Je zou kunnen verwachten dat de voorgaande specificatie een veel belangrijkere zaak is binnen een spaarzaam verbonden meta-analysis netwerk van een paar kleine proeven.

List of Publications

List of Publications

ARTICLES IN PEER-REVIEWED JOURNALS

Meta-analysis in small populations

Data generating models of dichotomous outcomes: Heterogeneity in simulation studies for a random-effects meta-analysis. **K Pateras**, S Nikolakopoulos, KCB Roes. [Statistics in Medicine], 2018

Interval estimation of the overall treatment effect in a meta-analysis of a few small studies with zero events. **K Pateras**, S Nikolakopoulos, D Mavridis, KCB Roes. [Contemporary Clinical Trials Communications], 2018

Prior distributions for variance parameters in a sparse-event meta-analysis of a few small trials. **K Pateras**, S Nikolakopoulos, KCB Roes. [Pharmaceutical statistics], 2020

Borrowing strength from short-term outcomes while accounting for sampling-based selection bias in orphan drug development processes. **K Pateras**, S Nikolakopoulos, KCB Roes. Under revision - [Currently in Statistics in Medicine]

Network meta-analysis

Participants' outcomes gone missing within a network of interventions: Bayesian modelling strategies. L Spineli, C Kalyvas and **K Pateras**. [Statistics in Medicine], 2019

List of Publications

Network meta-analysis and meta-analysis of a few trials

Association Between Blood Pressure Variability, Cardiovascular Disease And Mortality In Type 2 Diabetes: A Systematic Review And Meta-Analysis. M Chiriacò, **K Pateras**, A Virdis, M Charakida, D Kyriakopoulou, M Emdin, C Tsioufis, S Taddei, S Masi, G Georgiopoulos. [Diabetes, Obesity and Metabolism], 2019

A Bayesian meta-analysis on early tobacco exposure and vascular health. From childhood to early adulthood. G Georgiopoulos, D Oikonomou, **K Pateras**, S Masi, N Magkas, D Delialis, E Ajdini, V Vlachou, K Stamatelopoulos, and M Charakida. [European Journal of Preventive Cardiology], 2019

Antithrombotic treatment in cryptogenic stroke patients with patent foramen ovale: systematic review and meta-analysis. D Sagris, G Georgiopoulos, K Perlepe, **K Pateras**, E Korompoki, K Makaritsis, K Vemmos, H Milionis. [Stroke], 2019

Antithrombotic treatment in patients with stroke or transient ischemic attack and supracardiac atherosclerosis: systematic review and meta-analysis. D Sagris, G Georgiopoulos, I Leventis, **K Pateras**, LA Pearce, E Korompoki, K Makaritsis, K Vemmos, H Milionis, G Ntaios. [Neurology], 2020

Relative Efficacy of the Latest Therapeutic Options for Heart Failure with Reduced Ejection Fraction (Vericiguat, Sacubitril-Valsartan, Dapagliflozin): A Systematic Review and Network Meta-Analysis. A Aimo, **K Pateras**, K. Stamatelopoulos, CM Lombardi, M Senni, M Metra, C Passino, M Emdin, G Georgiopoulos. Submitted

Proprotein Convertase Subtilisisn-Kexin Type 9 inhibitors and stroke prevention: systematic review and meta-analysis. D Sagris, G Ntaios, G Georgiopoulos, **K Pateras**, E Korompoki, K Makaritsis, K Vemmos, H Milionis. Submitted

Device-detected atrial high-rate episodes duration and risk of thromboembolic events: systematic review and meta-analysis. D Sagris, G Georgiopoulos, **K Pateras**, G Ntaios. Submitted

Rare diseases

Applicability and added value of novel methods to improve drug development in rare diseases. M Mitroiu, K O Rengerink, C Pontes, A Sancho, R Vives, S Pesiou, M Fontanet, F Torres, S Nikolakopoulos, **K Pateras**, G Rosenkranz, M Posch, S Urach, R Ristl, A Koch, S Loukia, J H van der Lee, K CB Roes. [Orphanet Journal of rare diseases], 2018

Social sciences

Violence against women and unemployment a multidisciplinary and critical approach. K. Sklavou†, M. Valasaki†, **K Pateras**†, A Mastroyiannakis. [Dialogues in Clinical Neuroscience & Mental Health], 2020
† Equally contributed.

Έμφυλες διακρίσεις, βία κατά των γυναικών και εργασία: αναδυόμενες ανάγκες και κριτικές προσεγγίσεις. Μ Βαλασάκη, Κ Πατέρας, Κ Σκλάβου, Α Μαστρογιαννάκης. [Βία κατά των Γυναικών – Έμφυλη βία, Panteion University of Social and Political Sciences], 2020

Paediatrics

Association between fat mass through adolescence and arterial stiffness: a population-based study from The Avon Longitudinal Study of Parents and Children (ALSPAC). F Dangardt, M Charakida, G Georgiopoulos, S Chiesa, A Rapala, K Wade, A Hughes, N Timpson, **K Pateras**, N Finer, N Sattar, G Davey-Smith, D Lawlor, J Deanfield. [The Lancet Child & Adolescent Health], 2019

Birthweight by Gestational age reference centile charts for Greek neonates. A Tsagkari †, **K Pateras**†, D Ladopoulou, E Kornarou, N Vlachadis. Submitted
† Equally contributed.

Miscellaneous

Amyloid-β (1-40) and mortality in patients with non-ST elevation acute coronary syndrome: a cohort study. K Stamatelopoulos M Mueller-Hennessen, G Georgiopoulos, M Sachse, J

Boeddinghaus, K Sopova, A Gatsiou, C Amrhein, M Biener, M Vafaie, F Athanasouli, D Stakos, **K Pateras**, R Twerenbold, P Badertscher, T Nestelberger, S Dimmeler, HA Katus, A M Zeiher, C Mueller, E Giannitsis, K Stellos. [Annals of Internal Medicine], 2018

The complex relationship between serum uric acid, endothelial function and small vessel remodelling in humans. S Masi, G Georgopoulos, G Alexopoulos, **K Pateras**, J Rosada, G Serravalle, CD Ciuceis, C Borghi, S Taddei, G Grassi, D Rizzoni, A Virdis and the Study Group on Micro- and Macro-circulation of the Italian Society of Hypertension (SIIA). [Journal of Clinical Medicine], 2020

Wenckebach cycle length for the incidence of AV block in AVNRT patients treated with radiofrequency catheter ablation. S Chatzidou, C Kontogiannis, G Georgiopoulos, M Kosmopoulos, **K Pateras**, M Spartalis, K Stamatelopoulos, S Rokas. Submitted

REPORTS

Bayesian evidence synthesis for combining results of series of a few small trials. **K Pateras**, L Spineli, KCB Roes. [Revised Internal Asterix report], 2018

A review of frequentist methods for combining results of series of trials. I van der Tweel, **K Pateras**, GCM van Baal, KCB Roes. [Revised Internal Asterix report], 2015

CONFERENCES ABSTRACTS

Oral presentations [OP] - Poster presentations [PP]

[PP] Identifying and eliminating implausible gestational age birth weights: Greek female birth cohort 2011-2017. 41th Annual Conference of the International Society for Clinical Biostatistics, *July 2020 Krakow, Poland.*

[PP] Caution required in a poorly connected networks of interventions. 40th Annual Conference of the International Society for Clinical Biostatistics, *July 2019 Leuven, Belgium.*

[OP] The enlightening journey of three Data Generating Models. 31th Panhellenic Conference in Statistics, *May 2018, Lamia, Greece.*

[PP] The enlightening journey of three Data Generating Models: Heterogeneity in simulation studies for a random-effects meta-analysis. 38th Annual Conference of the International Society for Clinical Biostatistics, *July 2017, Vigo, Spain.*

[PP] Sparse-events evidence synthesis in small populations. 9th EMR and Italian Region of IBS conference, *May 2017, Thessaloniki, Greece.*

[OP] Heterogeneity under a sparse-events meta-analysis in small populations. 37th Annual Conference of the International Society for Clinical Biostatistics, *August 2016, Birmingham, United Kingdom.*

[OP] Strategies for dealing with heterogeneity between studies in rare diseases. 36th Annual Conference of the International Society for Clinical Biostatistics, *August 2015, Utrecht, The Netherlands.*

[PP] A review on methods for combining results of a series of trials. Clinical trials in small populations - Royal statistical society, *December 2015 X, London, United Kingdom.*

Acknowledgements

Acknowledgements

Prof. dr. Kit Roes, dear Kit, passing through all the obstacles that were appearing or I was seeing in front of me, your constant guidance and support was one of the main reasons that this thesis has been completed. Especially during 2015, when my levels of self-confidence were running low, you gave me the necessary time to come back. Thank you for that, and for the systematic long-distance support during the final years of this PhD.

dr. Stavros Nikolakopoulos, dear Stavro, αγαπητέ Σταύρο, unquestionably the impact that you had on my work efficiency cannot be described in a few sentences. This thesis would not have finished without your critical comments. Thank you for the mostly rainy but also sunny weekly meetings and all the discussions on statistical and non-statistical matters.

dr. Dimitris Mavridis, dear Dimitri, αγαπητέ Δημήτρη, your actual contribution in this thesis cannot be quantified in just one chapter. Your directions and guidance were brought to me at a period that I was fearing stepping back from this project. Thank you for showing patience to my initially long e-mails and unstructured draft manuscripts. I hope we may collaborate again in the near future.

dr. Victor Jong , dear Victor, dear ex room-mate, although your name appears only here in text and our methodological topics were different, this thesis contains small but multiple contributions from you. The high performance cluster tips and tricks, which you generously shared with me, aided to great extent my working efficiency. I wish the best for you and your family I am sure that you will fulfil your future ambitions.

Dear members of the assessment committee, Prof. dr. Armin Koch, dr. Caroline van Baal, dr. Irene Klugkist, Prof. dr. Leonard van den Berg, Prof. dr. Rene Eijkemans, dr. Saskia le Cessie, thank you for the time you devoted to study and evaluate this thesis.

Dear members of the mock committee, Fotis Polydoros, dr. Giorgos Georgiopoulos, Marian Mitroiu, dr. Marijn Hazelbag, dr. Miranta Antoniou, dr. Romin Pajouheshnia, dr. Ruben van Eijk and dr. Rutger van den Bor, thank you for voluntering to be one of my mock opponents.

Dear Caroline, Ingeborg, Kit and Rene, I would like to thank you all for interviewing, selecting and placing your trust in me back in November 2013 to conduct research on rare diseases within the Asterix project.

Acknowledgements

Dear Biostatistics and Research Support department colleagues (Bert, Caroline, Cas, Rebecca, Rene, Paul, Peter, all previous and current team members). Thank you for welcoming me within the group, thank you for the coffee breaks, cookies, cakes, thank you for the constructive feedbacks during our research meetings. I would like to especially thank those who were kindly there whenever I needed someone to share my thoughts and concerns.

To our department's team of multicultural mostly former but also current PhD students. Julien, Marian, Marteen, Putri, Rik, Rutger, Stavro, Victor, your presence in the department created an additional feeling of belonging. I cherish and I am grateful for each and every interaction we had.

I will never forget my first working room in the Netherlands (Str. 6.119) and of course all the people that shared a working day with me there. Madelief, Xin (Cindy), Jaike, Victor, Putri, Veerle, Sara, Anne Meike, Eline, Min and Michelle, I hope I gave back at least a small part of the positive energy you were sharing with me so kindly. All the best, in your current and future endeavours.

Being a member of the Asterix consortium gave me the opportunity to interact with people of diverse backgrounds. Without a doubt, my understanding of biostatistics and clinical trials was positively impacted by this interaction. Armin, Caridad, Caroline, Charlotte, Egbert, Eva, Ferran, Hanneke, Kit, Kristina, Lukas, Marian, Marleen, Martin, Katrien, Robin, Roser, Stavro, Stella and Susanne, thank you all for the feedback, sharing of knowledge and all the fruitful discussions.

My sincere appreciation to dr. Domenique Zomer and the cystic fibrosis group at the UMCU that gave me the opportunity to interact with their group and apply newly gained knowledge from the Asterix project.

To the High Performance Computing Cluster team of the University Medical Center Utrecht. Thank you all for the constant support and patience from my initial inefficient job submissions to the hopefully more efficient final ones.

To the people of CMT Prooptiki, Artemis, Caroline, Deppy, Dessy, Domna, Eirini, Giorgo B, Giorgo S, Ilia, Maria P, Maria V, Niko, Penny, Taso, thank you all for open-heartily providing

me with a steady environment to practice my skills and to participate in many European and national projects both in the health and social sector.

dr. Giorgos Georgiopoulos, αγαπητέ Γιώργο, thank you for the chance to practice my previous or recently acquired specific PhD skills in the field of cardiovascular research, a collaboration that has already led to numerous published and upcoming articles within national and international consortiums.

Prof. dr. Ioannis Ntzoufras, dear Ioanni, δάσκαλε, these words are here due to your belief in me and constant support. I am both fortunate and grateful to have you as a teacher.

To my first working group in the Hellenic Centre for Diseases Control and Prevention in the department of National archives (Public health, cancer and rare diseases), Fivo, Lia, Maria, thank you for inspiring me to perform research in the rare diseases area and for the great and fruitful collaboration.

Deciding after 29 years in Greece to pursuit a postgraduate degree in a foreign country would have been almost impossible without the daily interactions with you. Anna Maria, Anna, Asyer, Delphine, Eleonora, Federico, Frank, Katerina, Matevz, Matteo, Niko, Romin, Ruben, Sylvia and Tessa, my integration in the Netherlands would have been a lot harder without you. Matteo and Tessa another big thank you for helping in the translation of the summary in Dutch!

Teachers and fellow athletes from the Shudokan Kendo Utrecht training centre, thank you for educating me on devotion, focus, patience and for chasing up my fitness levels.

To all who shared walks and abstract conservations with me wondering among others if we are one that consists of many or many that make one, thank you.

Finally, the people that I left behind but never stopped supporting me through these years; Argiri, Christina, Elisavet, Giorgo, Ioanni, Manoli 1, Manoli (1), Niki, Sotiri, Vivi, you are my proof that distant friendships last. Elsa and Dimitri after this pandemic, just one request, let's make a deal and escape this war of mine. Miranda and Foti let's make five the countries that we meet each other. Among all of you additional praise to Manolia for visiting us the most in the Netherlands (7+1 times) and being there when we needed them the most!

Acknowledgements

To my two paranymphs, Matevž Rumpet and Sotiris Tsatsarounos. Matevž and Σωτήρη, except for here, you were both there for me in peculiar times and places. Hvala, σας ευχαριστώ.

Γονίδια, αδερφέ, θείοι, ξαδέρφια και παππούδια, σας ευχαριστώ και θα σας ευχαριστώ για κάθε βήμα της ζωής μου. Χωρίς το δομημένο και ασφαλές περιβάλλον που γενναιόδωρα μου παρείχατε, το βιβλίο αυτό δε θα υπήρχε.

Acknowledgements

My dear Mary, without your constant support, infinite patience and meaningful contribution (textual reviewing), this journey will not have started, will not have continued and would not have taken this path. Thank you for the last 14 years and for reminding me that *"We don't have time to be timid. We must be bold and daring"*[‡].

Αγαπημένη μου Μαίρη, σε ευχαριστώ για τα τελευταία 14 χρόνια. Χωρίς τη στήριξή, την υπομονή, την επιμονή και την ουσιαστική συνδρομή σου, η διαδρομή αυτή δε θα ξεκινούσε, δε θα συνέχιζε και φυσικά δε θα τελειώνε ποτέ με αυτόν τον τρόπο. Ελπίζω να μπορέσω να στο ανταποδώσω κάποτε. Α, και κάτι για σένα, ϑ(σΛ)γΠ.

[‡]Lumière

Προς όλους, μεθύστε.
Μεθύστε χωρίς διακοπή.
Με κρασί,
με ποίηση,
ή με αρετή.
Όπως σας αρέσει.
Αλλά μεθύστε!

Σὰρλ Μπωντλαίρ
(Μεθύστε, Μελοποίηση - Διάφανα Κρίνα)

To everyone, be drunk.
Be continually drunk.
With wine,
with poetry,
or with virtue.
As you choose.
But be drunk.

Charles Baudelaire
(Be Drunk X)

About the author

About the author

Konstantinos Pateras was born in Athens, Greece, on the 28^{th} of March, 1985. He studied at the Athens University of Economic and Business - AUEB - and the National and Kapodistrian University of Athens - UOA from which he obtained a bachelor's degree in Statistics, as well as a master's degree in Biostatistics. His master thesis in UOA was performed under the supervision of Ioannis Ntzoufras on Bayesian variable selection in Normal and Binomial models with application in Medical research.

His first formal contact with Biostatistics occurred in 2006 during his one year collaboration with the 2^{nd} University Psychiatric Clinic, "*Attikon*" General University Hospital. Between 2011-2013, he became affiliated with the Hellenic Center of Disease and Control Prevention where he focused on resetting the National Public health registry (ERDF 2007-2013) and providing guidance on the Cancer and Rare diseases registries. Shortly after this he joined the Julius Center, University Medical Center Utrecht - UMCU, in order to conduct research on evidence synthesis in rare diseases for the Asterix project under the supervision of Kit Roes (EU-FP7 2013-2017). Among his other interests was the partial development of gamlss.demo an R package for demonstrating a variety of distributions for educational purposes, as well as the development and technical administration of Grstats forum "*Leshi Filon Statistikis*", a platform that informs Greek statisticians and supports non-statisticians. He was a founding member of STArT Thinking, a voluntary initiative that motivated undergraduate and postgraduate statisticians in Greece about the value and practical applications of statistics.

He cherished pro bono scientific collaborations with the University of Hannover, Department of Biometrics on (Network) meta-analysis. He contributed in NEMO (GRF 2017-2019), Cardiovascular projects (EACVI 2019-2020, EFSD 2020-2021) and at least 10 projects on social inequalities or public health (ERDF 2007-2013, EU 2014-2020). He also provides statistical consultancy to CMT Prooptiki Ltd, the General University Hospital Alexandra, the Department of Clinical Therapeutics, National and Kapodistrian University of Athens and the School of Biomedical Engineering and Imaging Sciences, King's College London.

Making the most of a few small population clinical trials

Konstantinos Pateras