Extensions and applications of generalized linear mixed models for network meta-analysis of randomized controlled trials

Dissertation to obtain the doctoral degree of Agricultural Sciences (Dr. sc. agr.)

Faculty of Agricultural Sciences

University of Hohenheim

Biostatistics
Institute of Crop Science

submitted by

Anna Wiksten

Born in Siuntio, Finland

2022

This thesis was accepted as a doctoral dissertation by the Faculty of Agricultural Sciences at the University of Hohenheim on 02 November 2022.

Date of oral examination: 08 December 2022

Examination Committee

Head of the Examination Committee: Prof. Dr. Jörn Bennewitz

Supervisor and Reviewer: Prof. Dr. Hans-Peter Piepho

Co-Reviewer: Prof. Dr. Ekkehard Glimm

Additional Examiner: Prof. Dr. Christine Wieck

Contents

Acknowledgements

Kiitos - Thank You:

Hans-Peter Piepho – for being a true role model of a scientist and a leader. I was always impressed by your science driven, collaborative, meaningful, pragmatic, and fast feedback on the manuscripts and thesis. It has been a pleasure to work with you.

Ekkehard Glimm –for working on the topic with me from the beginning until the end and for always ensuring that my maths are correct, both for this thesis and my daily work at Novartis. I admire your skills in solving complicated, but still relevant, problems with pen and paper. Thank you for your friendship too and for all the good discussions on topics beyond statistics.

Sandro Gsteiger, Gerta Rücker, and Guido Schwarzer – for giving me the opportunity to work with you on applications and work together on the manuscripts included in the thesis. I learned a lot from all of you and had fun too.

Sini Suokas – for your friendship and encouragement on critical moments.

Elina Asikanius – for your friendship and for sharing me all your knowledge and rational thinking about drug development and statistics on our walks in Basel and Alps.

Pasi Korhonen – for showing me an example of combining excellence in statistics and drug development and providing me opportunity to work on projects which contributed to this thesis.

Kapil Sen and Insa Gathmann – for your leadership as my managers, encouraging me to finalize the thesis beside my work at Novartis, and giving me the opportunity to take extra weeks off for it.

All Mediasres Early Stage Researchers and my colleagues at Novartis methodology group – for the all the collaboration and interactions during 2012-2015, when the grounds for this PhD were set.

Markus Lange – for being a great colleague and friend for the 3 years when we shared the office at Novartis. It is valuable and fun to have colleague with whom no topic is a taboo for discussion.

Hoffman-La Roche and Novartis Pharma AG – for providing me inspiring real-life applications problems to do research on.

Finland – for providing me education from elementary school up to MSc, which enabled me to do research in statistics.

Terhi Katiskalahti, Pekka Heikkilä, Sanna Ström, and Tuomas Kemppainen - for your friendship and relaxing times and reminding me to keep my feet on the ground. And for mentoring and onboarding me in my first job in drug development.

Hanna Susi – for being such a good friend, cousin, and peer support when it was difficult in PhD work.

My sister Iida Korpela and all my sisters-in-law, brothers-in-law – thank you for all the discussions on maths, medicine, social welfare, writing, and any other topics over the years.

Jussi Nyman – for letting me build miniature aircraft with you and taking me fishing with you on early mornings.

Pirjo and Raimo Wiksten – for your love and support.

Meki(my dog) – bow wow!

Saija Siivonen, Juha Javanainen, and Tiina Kirsilä – thank you for everything. I feel privileged to have 3 best friends.

Saija, thank you for letting me sit next to you at school when my family moved to Turku in 1996.

Juha, thank you for asking me to dance with you in 2000.

Tiina, thank you for choosing to come to Turku instead of Rovaniemi in 2005.

Dedication

This thesis is dedicated to all my peers.

“ Maailma, jossa on loputtomasti dataa, tarvitaan just niitä, jotka osaa tehdä sillä tiedolla jotain yhteiskunnallisesti tai lääketieteellisesti hyödyllistä ja merkittävää.”

and the same in English

“A world with limitless data needs people with the ability to turn it into impactful insights for medicine and society”

– Heini Alsio,

a journalist and a peer of mine

1 General introduction

1.1 Analysis of groups of experiments in agricultural research and meta-analysis

In agricultural research the same type of experiment for a treatment factor of interest is often repeated in different locations, over a number of years or different labs for example. This type of research requires methods for combining results from several experiments and analyzing them jointly. Methods for analyzing groups of experiments jointly were first published in a paper by Yates and Cochran in 1938 (Yates and Cochran, 1938). This paper set the stage for methods later developed for meta-analysis and for the combination of estimates from different experiments.

Meta-analysis is a general term for a method which has later been used in several areas of research where quantitative experiments are performed to make inference about different factors/interventions. So far the most common area for applying meta-analysis and network meta-analysis for evidence synthesis has been medicine but these methods are popular in other disciplines as well for example psychology (Curran and Hussong, 2009), ecology (Koricheva et al., 2013) or agriculture (Madden et al., 2016).

The term network-meta-analysis has not been used much in agricultural research, but it has increasing interest there as well and some methods developed in context of medical research have also been applied in in agricultural research (Cordova et al., 2017; Machado et al., 2017; Paul et al., 2019; Sauer et al., 2008).

This thesis will explore and develop evidence synthesis methods in applications to randomized clinical trials.

1.2 Meta-analysis and network meta-analysis of randomized clinical trials

In medical research an effect of a treatment or medical intervention is often studied in a randomized clinical trial. When the same treatment is studied in several trials the need for combining results and making unified evidence synthesis arises. In this thesis we will use either the term "*trial*" or "*study*" for the repeated experiment and the factor of interest is called "*treatment*".

In pharmaceutical research it is often necessary to make inferences on more than two treatments and their relative efficacy. This inference can be used to support decision making in drug development programs, re-imbursement decision of payers and even physicians to make decisions: Which treatment is the best choice for my patient?

The term meta-analysis is usually used when only two treatments have been compared in several trials. The need to compare several treatments leads to an extension of meta-analysis and is called network meta-analysis (NMA). In network meta-analysis the included trials may have more than two treatments analyzed and the combination of treatments can be different in different trials. In network meta-analysis several treatments from several trials are analyzed jointly to provide a comparison of all treatments of interest. Figure 1 shows an example of indirect and direct comparisons, displayed as a graph, in which the nodes are used to illustrate the different treatments and the edges present if the two treatments have been compared in the same randomized trial. The treatments form a network which might be connected through direct (treatments compared in same study) or indirect (in different studies through same comparator) comparisons. Often the network consists of both direct and indirect comparisons for different treatments. For example there might be one trial comparing *A* and *B* directly and two other trials; one where *A* is compared with *C* and other where *B* is compared with *C*, and through *C* an indirect comparison of *A* and *B* is possible. The NMA estimate for the comparison *A* and *B* is a combination of direct and indirect comparisons.

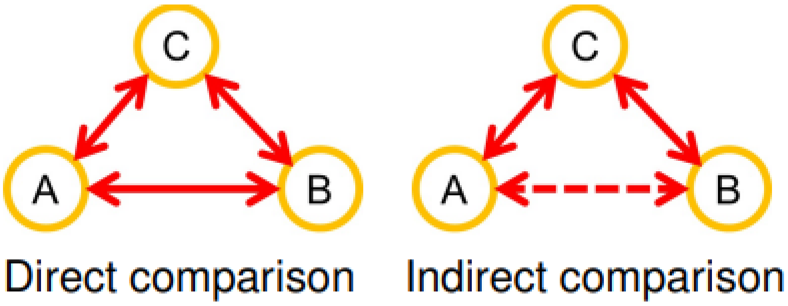


Figure 1 Direct and indirect comparison of treatment A and B. Solid lines present a direct comparison (the treatments A and B have been compared in a same trial). Dashed line presents an indirect comparison (treatments A and B have been compared with C in separate trials, and they can be compared (indirectly) trough common comparator (C))

Having 3 treatments which are compared in two separate trials is the smallest possible network. Usually the network of trials includes more treatments and more trials. Figure 2 illustrates how extensive a network of studies can be; it is the network of data used for the application presented in the Chapter 3 of this thesis. The data had over 50 treatments compared in over 200 trials.

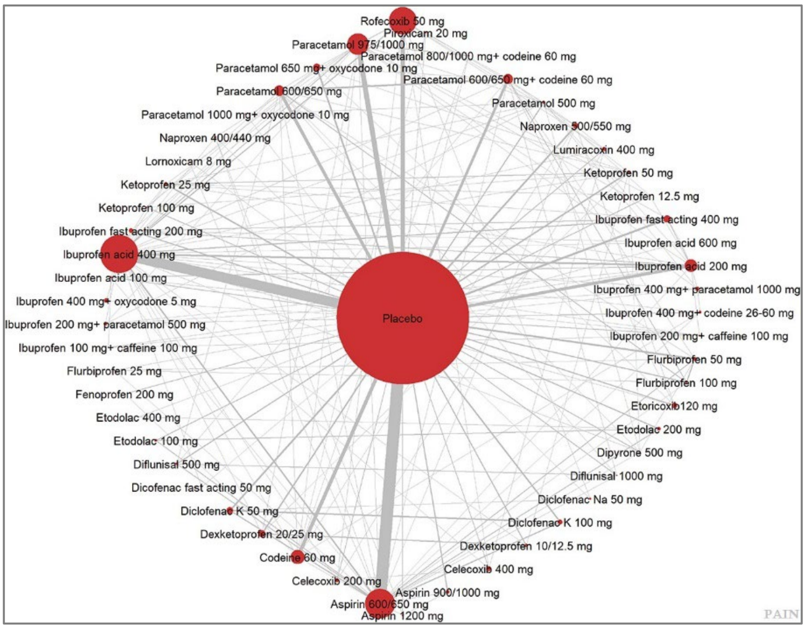


Figure 2 Network diagram for the pain relief network meta-analysis

The literature on network meta-analysis has largely been based on so-called baseline contrast model (Lu and Ades, 2006). In the baseline contrast model one of the treatments (often the one which comes alphabetically first) is selected as “baseline” treatment and other treatments are modelled as contrasts to this treatment. As not all trials in a network may contain the same “baseline”-treatment, the baseline treatment may vary from trial to trial. One alternative approach is to use arm-based models as proposed by (Jones et al., 2011; Piepho et al., 2012). These types of models can be estimated using standard analysis-of-variance techniques. Having trial as fixed factor and treatment as fixed or random factor in the model is very similar to the baseline contrast model regarding the assumptions and numerical results from estimation. These arm-based models can be extended to different types of data structures and outcomes using the generalized linear models. This thesis focuses on using generalized linear mixed model theory in modelling and estimation for network meta-analysis of randomized clinical trials.

The arm-based network meta-analysis is sometimes criticized for not following the principle of concurrent control. This is the case for models (both in meta-analysis and in network meta-analysis) which don’t have any term for trial or if the trial main effect is specified as random (Senn, 2000). In this thesis, the models applied always include the main effect for trial as fixed and therefore the principle of concurrent control is followed similarly as in baseline contrast based models for network meta-analysis.

In some cases the arm-based and baseline contrast model result in the same model and/or exactly equal estimates. In fixed-effects network meta-analysis the arm-based modelling with trial and treatment as fixed factors the model can be easily converted into baseline contrast model and they can be seen just as different parameterizations of the same underlying model. In random-effects network meta-analysis the two approaches make a different assumption of the random effects distributions and the models are different. Even in this case they may yield same estimates for the treatment contrast parameters depending on the estimation technique and even if not exactly the same in many cases in very close agreement (Jones et al., 2011; Piepho et al., 2012; Wiksten et al., 2020).

1.3 Generalized linear models in network meta-analysis

In the ANOVA type of model with arm-based parametrization (Jones et al., 2011; Piepho et al., 2012) all network analysis models can be specified in a general form of generalized linear models by specifying the linear predictor, likelihood, and link function.

1.3.1 Fixed effect model

For the fixed effect model the linear predictor can be written as

$$\eta_{jk} = \alpha_j + \theta_k$$

where α_j stands for fixed study effect for trial *j* and θ_k is the fixed treatment effect for treatment *k* (*k*= 1,2,3,... *or k=A,B,C,D*...). In fixed effect models the assumption made is that the true treatment effect is the same across all trials and any differences seen between trial specific treatment effects is only due to random sampling. Even if the treatment effects are assumed to be the same there might be differences in other prognostic factors between trials and these often unknown prognostic factors (different location of trial, different patient populations, etc.) can be taken into account by adding fixed trial effects α_j. Having the trial effect α_j as fixed means the estimates of treatment contrasts $\theta_{k_2} - \theta_{k_1}$ are weighted averages, with weights defined by the size of the trial and taking into account the size of the trial. The fixed trial effect also ensures that treatment comparison are done within study and the model respects the randomization (Senn, 2000).

1.3.2 Random effects model

For the random effect model the linear predictor can be written as

$$\eta_{jk} = \alpha_j + \theta_k + u_{jk}$$

where

$$\boldsymbol{u_j} = \begin{pmatrix} u_{j1} \\ u_{j2} \\ \vdots \\ u_{jk} \end{pmatrix} \sim N_{a_j}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & \sigma_u^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_u^2 \end{pmatrix}\right)$$

where N_{a_j} denotes the multivariate normal distribution and a_j is the number of treatments in the jth trial. By adding the random effect u_{jk}, which is an interaction term, we make an assumption that the underlying true treatment effect may not be the same for all trials and we need to account for that heterogeneity. In network meta-analysis it is common to make the assumption of having a single variance parameter for all treatments (or treatment contrasts). Other types of covariance structures may be assumed as well. For example one could make an assumption that the between trial variance is heterogeneous for different treatments and this would yield to covariance matrix with a different variance parameters for each treatment and the variance parameters on the diagonal of the covariance matrix would be unique for each of the *k* treatments.

1.3.3 Relation to baseline contrast model/parametrization

When using fixed effect models the more common contrast-based parametrisation of NMA (Lu and Ades, 2006) and the arm-based NMA will result in equivalent inference for the treatment contrasts as the contrast-based formulation can be obtained by re-parameterization, namely

$$\eta_{jk} = \begin{cases} \mu_{jb}, & \text{if } k = b,\ b = A, B, C, \ldots \\ \mu_{jb} + d_{bk}, & \text{if } k \text{ alphabetically'after } b \end{cases}$$

where $\mu_{jb} = \alpha_j + \theta_b$ and $d_{bk} = \theta_k - \theta_b$. Of note, the baseline contrast model cannot be derived uniquely from arm-based model specification as the selection of baseline treatment is not unique.

For the random effect model the linear predictor for baseline contrast model can be specified as

$$\eta_{jk} = \begin{cases} \mu_{jb}, & \text{if } k = b,\ b = A, B, C, \ldots \\ \mu_{jb} + \delta_{bk}, & \text{if } k \text{ alphabetically'after } b \end{cases}$$

Where δ_{bk} is the study specific treatment effect for treatment *k* versus the trial specific baseline treatment *b*. The random effects $\boldsymbol{\delta_j}$ are assumed to be normally distributed as

$$\boldsymbol{\delta_j} = \begin{pmatrix} \delta_{j,2} \\ \delta_{j,3} \\ \vdots \\ \delta_{j,a_j} \end{pmatrix} \sim N_{a_j-1}\left(\begin{pmatrix} d_{t_{j1}t_{j2}} \\ d_{t_{j1}t_{j3}} \\ \vdots \\ d_{t_{j1}t_{ja_j}} \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2/2 & \cdots & \tau^2/2 \\ \tau^2/2 & \tau^2 & & \tau^2/2 \\ \vdots & & \ddots & \vdots \\ \tau^2/2 & \tau^2/2 & \cdots & \tau^2 \end{pmatrix}\right)$$

Of note here, t_{j1} refers to the trial-specific baseline treatment and may differ from trial to trials and $t_{j2},\ t_{j3}, \dots t_{ja_j}$ may be any combination of treatments *B, C, D…* depending on treatments included in the given trial and selected trial-specific baseline treatment.

For random effects models, the main difference between classical ANOVA-based parametrization and baseline-contrast parametrization is in the assumption of distribution of random effects. This is discussed in more detail in Chapter 2 for normally distributed response variable. For other type of responses it has been discussed in literature in (Jones et al., 2011; Piepho et al., 2018, 2012).

1.3.4 Model estimation

The model parameters for both types of models can be estimated in both Bayesian and frequentist setting. In case of using "non-informative" or vague priors in Bayesian analysis the numerical parameter estimate values will be in close agreement especially if the sample size is large enough.

1.4 Motivation and research objectives

The theoretical basis for using the generalized linear models for network meta-analysis with arm-based and frequentist approach was set in the publications by (Jones et al., 2011) and (Piepho et al., 2012). Both of the papers discussed the idea and presented theoretical justifications for the modelling approach with application in the binomial smoking cessation dataset from (Hasselblad, 1998). These two publications set the stage for applying widely used statistical methodology and knowledge originally applied mostly to agricultural research for network meta-analysis of clinical trials. However, despite the groundbreaking work of Jones et al. (2011) and Piepho et al. (2012), there has been a gap of applying and evaluating the usability of the methods in applications arising from urgent medical questions with different endpoints and more complicated study designs or analysis problems. The objectives of this thesis are to apply and develop the methods in different applications and provide a comprehensive summary of the benefits and limitations of the modelling framework and provide guidance for other researchers and analysts how to implement the methods for different network meta-analysis problems. This will be done through applications, which have arisen from real analysis problems in drug development.

The specific objectives to address some of the urgent research gaps are:

1. Explore and extend the methods in different applications
    - Different levels of aggregation and modelling treatment-by-covariate interactions – this is a very common problem in the pharmaceutical industry, where the companies have access to individual patient data from own studies, but only aggregated data from studies performed by other organizations.
    - Extensive networks with many treatments and trials – this research problem is motivated by actual evidence synthesis need for an extensive network arising from pain medications used and studied in over 200 clinical trials.
    - Different type of outcome data (time-to-event) with complicated modelling problem – this research problem has become very urgent when the new mode of action cancer therapies are compared with the older treatments and the assumption of proportional hazards over time may not hold anymore.
2. Explore the differences between baseline contrast parametrization and arm-based parametrization
    - The relationship between the two modelling approaches has been described in the original papers where the arm-based models were introduced, however using the arm-based methods for different real applications has been lacking and one objective of this thesis is to provide more case studies with comparison of the methods.
3. Explore methods of between study variance estimation in meta-analysis and network meta-analysis
    - Between study variance estimation is one of the most critical components of NMA and the aim of this thesis was to also evaluate the properties of between-study variance estimation for arm-based models, how it compares with baseline contrast model, and provide insights on preferred estimation methods for different applications.

4. Provide software implementation and example codes for a variety of models
    - The analyses in the thesis are performed using the standard generalized linear model functions and procedures from R or SAS.

1.5 Outline of the thesis

This thesis is a cumulative thesis and each chapter is based on an individual manuscript or journal article.

Chapters 2-4 explore and develop the methods for three different applications with different datatypes and scientific questions. All three manuscripts present different problems and have their own challenges making the existing NMA methods potentially complicated to be used, as will be outlined briefly further below:

- Chapter 2 tackles a modelling problem of two different aggregation levels
- Chapter 3 tackles a problem with extensive network with many trials and treatments
- Chapter 4 tackles a problem with complex modelling of underlying data

Chapter 2 General linear models for combining individual patient data and aggregated data in network meta-analysis

This manuscript develops the arm-based methods and models in situations where some of the trials provide patient level data and others aggregated data from publications. This is a common situation in pharmaceutical companies, where the statistical analysts usually have access to data for studies performed in-house, but only aggregated data for competitor studies. Similar modelling needs may also arise in agricultural sciences if a research team has access to original data of their own experiment but only aggregated published results from other experiments. The manuscript introduces meta-regression models to adjust for covariates and introduces an approach how to evaluate loss of precision due to aggregated data vs individual patient data.

Chapter 3 Estimating relative efficacy in acute postoperative pain: network meta-analysis is consistent with indirect comparison to placebo alone

This paper applies the method in a real life dataset with pain medications used in acute postoperative pain. The outcome of interest was binomial, whether a subject experienced pain relief or not. The dataset used for NMA included 261 trials with 52 different treatment and dose combinations, making it extraordinarily rich and large. The manuscript provides an example of analyzing extensive and large network with efficient standard SAS generalized mixed model procedures. We also provide different visualization to present the network and compare the treatments.

Chapter 4 Nonproportional Hazards in Network Meta-Analysis: Efficient Strategies for Model Building and Analysis

This manuscript develops the method for a case of time-to-event-outcome extracted from published Kaplan-Meier curves of survival analyses. This re-generated individual patient data was then used to model and compare survival functions and hazards of different treatments. This manuscript introduces the methodology for a new type of outcome, time to event and also introduces efficient and time saving approach to find the most suitable model for non-proportional hazards. It also compares the existing, baseline contrast model, methodology to arm-based modelling in a new type of problem and provides a comparison of the methods.

Chapter 5 Hartung–Knapp method is not always conservative compared with fixed-effect meta-analysis

This chapter explores different methods for between-study variance estimation in standard meta-analysis in a set of 157 meta-analyses from Cochrane database. Estimating between-study variance is one of the most important part of the evidence synthesis, firstly to evaluate the amount of between-trial variance and whether the fixed- or random effects model is applicable. Secondly, it is important to properly account for the potential between-trial variation in treatment effect estimation. Hartung-Knapp method has been one of the most popular methods in standard two-treatment meta-analysis. This paper evaluates the properties of the method in large set of trials with binomial outcome and makes recommendations for model checking and useful sensitivity analyses.

Chapter 6 provides general discussion and conclusions. Finally, the work is summarized in Chapter 7.

1.6 List of included manuscripts

Wiksten, Anna, Glimm, Ekkehard. General linear models for combining individual patient data and aggregated data in network meta-analysis. Draft manuscript.

Moore, R.A., Derry, S., Wiffen, P.J., Banerjee, S., Karan, R., Glimm, E., Wiksten, A., Aldington, D., Eccleston, C., 2018. Estimating relative efficacy in acute postoperative pain: network meta-analysis is consistent with indirect comparison to placebo alone. PAIN 159, 2234–2244. https://doi.org/10.1097/j.pain.0000000000001322

Wiksten, A., Hawkins, N., Piepho, H.-P., Gsteiger, S., 2020. Nonproportional Hazards in Network Meta-Analysis: Efficient Strategies for Model Building and Analysis. Value in Health 23, 918–927. https://doi.org/10.1016/j.jval.2020.03.010

Wiksten, A., Rücker, G., Schwarzer, G., 2016. Hartung–Knapp method is not always conservative compared with fixed-effect meta-analysis. Statistics in Medicine 35, 2503–2515. https://doi.org/10.1002/sim.6879

2 General linear models for combining individual patient data and aggregated data in network meta-analysis

General linear models for combining individual patient data and aggregated data in network meta-analysis

Anna Wiksten and Ekkehard Glimm

Abstract

Network meta-analysis (NMA) is a method where multiple treatments from several trials are compared in a single analysis. Pharmaceutical companies perform NMAs to support submissions of new drugs or for market access and reimbursement purposes. Furthermore, in the development of a new drug, NMA may be used to support planning of future studies. When NMA is performed within a company, the company has access to individual patient data (IPD) of their own studies whereas usually only the aggregated data(AD) is available from the competitor's studies. In this presentation, we will introduce different models to perform NMA with IPD and AD and the methods will be applied to a case study. Both frequentist and Bayesian approaches will be considered. We will also present different NMA models for analyzing data with an individual-patient-component that includes additional covariates with a potential interaction to the treatment effect. Results from a simulation study comparing performance of the different models and different types of data will be presented.

1 Introduction

Network meta-analysis(NMA) is a method were more than two treatments from several clinical trials are compared in a single analysis. The NMAs often used for decision making by regulatory and government authorities. The pharmaceutical companies may perform a NMA to support submission of new drug and also for market access and reimbursement purposes. Also in the development phase NMAs can be used when making decisions for example continuing from phase 2 to phase 3[1] or planning future studies based on the evidence from NMA.

NMA may be performed either on aggregated data(AD) level or individual patient data(IPD) level. Generally IPD level analyses are considered as a gold standard in meta-analysis [13]. When the pharmaceutical companies perform a NMA inhouse they have of course access to IPD of their own studies as from competttitor studies usually only the AD is available.

There is a extensive literature on the network meta-analysis of aggregated data. The Bayesian methods are very popular in the field of NMA, but frequentist methods have been proposed as well[4, 7]. Often bayesian methods are used with vague priors and the results are very similar with frequentist analyses. The advantages of bayesian methods are clear when we have some prior information from some of the treatments from studies which are not included in the NMA and we want to incorporate the information in our NMA as prioirs. On the other hand when no prior information is available it might be difficult to find truly un-informative priors.

One important question for decision making is which treatment works best for which patient population and if there are differencies between treatment effects in different patient groups. For categorical covariates, e.g. sex, desease severity etc subgroup analyses may be used to analyse the treatment differences in different subpopulations. To be able to perform subgroup analyses the original studies have to report the results by subgroups or if the studies different inclusion criteria the overall results can be used. If the studies do not report the results by subgroups the percentage of categorical covariate value may be used as continuous covariate in metaregression. Meta-regression may be used to examine the effect of continuous covariates on the treatment effect, however the meta-regression with AD suffers from lack of power[5] and is subject ecological bias. When performing metaregression with aggregated covariate values special caution should be used.

When the IPD is available the treatment-by-covariate interactions can be modelled on patient level and the analyses have greater power to detect potential interactions. Achieving IPD for all of the studies in NMA is usually very difficult or even impossible. The studies included in the NMA are usually performed by different companies and all parties may not be willing to share their data. Recently there has been a growing demand for pharmaceutical companies to publish all clinical trial data.

There are several papers presenting the combining of IPD and AD in traditional two treatment meta-analysis[14, 9]. On NMA there are some recent papers about combining IPD and AD by Jansen[3] and Donegan et. all [2]. Both of these papers are based on the so-called baseline contrast model[6] and they are fitted in Bayesian framework.

The aim of this article is to extend the NMA methodology based on ANOVA framework to setting where some of the studies provide IPD and some of the studies only AD and to extend the models to analyse treatment-by-covariate interactions. We will apply the proposed methods in to a motivating example dataset and compare the results with existing methodology. We will also evaluate the performance of different models trough a simulation study.

The outline of the article is as follows. We first introduce the motivating dataset in Section 2. In Section 3 we describe our proposed models and the existing models and apply the methods in the example dataset. In Section 4 we describe the simulation study and the results from simulations.We conclude wit discussion in Section 5.

2 Motivating example

For confidentiality reasons the data used in example is simulated, but the simulations are based on real data. We are interested in comparing the efficacy of a treatment A with treatment B in a network meta-analysis and we have IPD available for all in-house studies and aggregated data for other studies. The primary efficacy outcome is a continuous variable and it is assumed to be normally distributed. There is a potential continuos treatment effect-modifying-covariate and we are want to model the interaction between the treatment and covariate. The aggregated studies are reporting study summaries per treatment arm and mean covariate value per treatment arm. Table 1 shows the treatments and sample sizes per each study.

3 Methods for estimating the treatment effects in network meta-analysis combining IPD and AD

In this section we will present methods for performing NMA combining IPD and AD. In all situations we are considering continuous outcome variable and in the models with effect-modifying-covariate the covariate is continuous and we will consider both frequentist and bayesian methods for NMA.

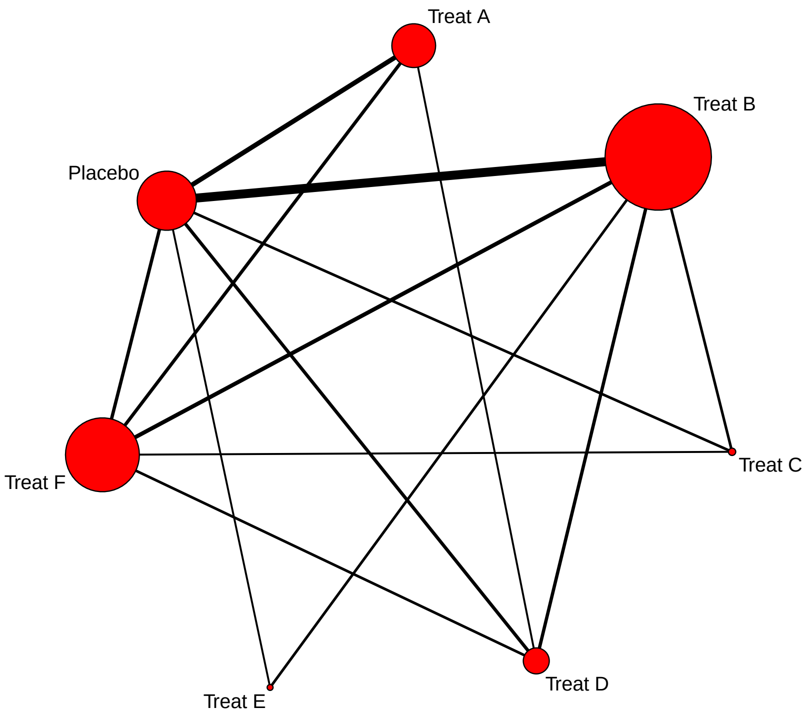


Figure 1: Network diagram of the studies included in the meta-analysis. The nodes represent the treatments in NMA and the straight lines represent the direct comparisons between two treatments. The size of the nodes is proportional to the number of subjects in each treatment arm

3 Models without additional treatment-by-covariate interaction

3 Two-way linear model for IPD and AD

Let us first consider situation where we have individual patient data available for all studies. Now for the IPD network meta-analysis we can use two-way linear mixed model as proposed by [4] and [7]. The two-way linear mixed model for IPD is

$$\begin{aligned} y_{ijk} &= \alpha_j + \theta_k + u_{jk} + \epsilon_{ijk} \\ u_{jk} &\sim N(0, \sigma_u^2) \\ \epsilon_{ijk} &\sim N(0, \sigma_j^2) \end{aligned} \tag{1}$$

where y_{ijk} is the response for ith subject in jth study with treatment k, α_j is the fixed study effect, θ_k is the fixed treatment effect, and u_{jk} is the random treatment-by-study interaction term in study j for treatment k and it is assumed to be normally distributed with mean 0 and between study variance σ_u^2, and ϵ_{ijk} is the residual error term for observation y_{ijk} and σ_j^2 is the variance of the residual errors in study j . The random term u_{jk}^2 and the residual error term ϵ_{ijk} are assumed to be stochastically independent. In model 1 we need a parameter condition for one of the fixed effects. Since we are typically interested

	Study	A	B	C	D	E	F	Plac	x mean	x sd
AD	1	0	200	0	0	0	0	100	0.99	0.42
studies	2	0	100	0	0	50	0	0	1.02	0.44
	3	0	360	0	0	0	0	180	0.86	0.37
	4	0	854	427	0	0	0	425	1.06	0.44
	5	0	1028	45	0	0	542	419	1.07	0.50
	6	0	330	0	0	327	0	334	1.07	0.48
	7	0	211	0	0	0	0	205	1.09	0.56
	8	0	1707	0	0	0	1714	0	0.87	0.3
	9	0	760	0	0	0	776	0	1.09	0.43
	10	0	161	0	0	0	0	155	1.02	0.48
	11	0	150	0	0	0	0	155	1.1	0.5
	12	0	473	0	471	0	478	229	1.04	0.48
	13	0	260	0	258	0	0	260	1.05	0.46
	14	0	251	0	249	0	0	247	0.95	0.48
IPD	15	523	0	0	0	0	266	267	1.11	0.50
studies	16	550	0	0	0	0	0	265	1.06	0.46
	17	301	0	0	0	0	0	154	0.87	0.37
	18	325	0	0	0	0	326	0	1.04	0.49
	19	222	0	0	0	0	0	218	1.06	0.50
	20	216	0	0	0	0	0	215	1.04	0.47
	21	702	0	0	707	0	689	0	0.85	0.30
	All	2839	6896	472	1685	383	4791	3835	1.00	0.44

Table 1: Number of observations in each study

in the differences of the treatments it is a natural choice to put the parameter condition on θ. If we set θ_K equal to zero, then all other parameters θ_1 to θ_{K-1} represent the treatment difference comparen to the last treatment and the fixed study effects α_j represents the the expected response for treatment K in study j. No parameter condition is needed for the random effects u_{jk}.

Let us assume that we have only aggregated data available from the studies included in network meta-analysis. Usually invididual studies report summaries per treatment arm per study. In case of a normally distributed continuous outcome the least square mean estimate of the response and its standard error from each treatment arm in each study are typically reported. If no additional covariates are included in the analysis of the individual studies the least square mean estimate is the sample mean and standard error is the square root of the residual error variance divided by the square root of sample size per treatment arm. Let us denote the least square mean for treatment k in study j as $y_{\cdot jk}$. Now the individual patient data model (1) for aggregated data can be written as

$$\begin{aligned} \bar{y}_{\cdot jk} = \frac{\sum_{i=1}^{n_{jk}} y_{ijk}}{n_{jk}} &= \frac{\sum_{i=1}^{n_{jk}}(\alpha_j + \theta_k + u_{jk} + \epsilon_{ijk})}{n_{jk}} \\ &= \alpha_j + \theta_k + u_{jk} + \bar{\epsilon}_{\cdot jk} \qquad (2) \\ & u_{jk} \sim N(0, \sigma_u^2), \bar{\epsilon}_{\cdot jk} \sim N(0, \frac{\sigma_j^2}{n_{jk}}) \end{aligned}$$

Since we observe only one observation per treatment and study, the estimation of the σ_j^2 is not possible from the aggregated data. The estimated value of σ_j^2 is treated as known parameter and we have

$$\bar{\epsilon}_{\cdot jk} \sim N(0, \frac{\hat{\sigma}_j^2}{n_{jk}})$$

Since $\hat{\sigma}_j^2$ is considered fixed, it is irrelevant how it was derived. In an ordinary one-way ANOVA, it would be estimated as

$$\hat{\sigma}_j^2 = \frac{(n_{j1}-1)\hat{\sigma}_{j1}^2 + ... + (n_{jK_j}-1)\hat{\sigma}_{jK_j}^2}{n_j - K_j}$$

where K_j is the number of treatments in study j and

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^{n_{jk}}(y_{ijk} - \bar{y}_{\cdot jk})^2}{n_{jk}-1}$$

is the residual variance estimate in study j for treatment k.

If we have individual patient data available for some of the studies and aggregated data for the other studies the model for IPD studies takes the form 1 and for AD studies the form 2, and both types of data will contribute to the estimation of the shared parameters $\theta_1, ..., \theta_k$ and σ_u^2.

In the model 1 the treatment difference A vs B is $\theta_1 - \theta_2$. The fixed sudy effect in these models means that the between-study information on treatment effect is not recovered and the model respects the principle of concurrent control[11]. Following the principle of concurrent conrol means that treatment effects are only judged by within study comparisons at the same time and hence randomisation is reserved.

3 Baseline contrast model

A popular model in the field of network meta-analysis is a model based on baseline contrasts intoduced by Lu and Ades[6]. The baseline contrast model for IPD is

$$y_{ijk} = \mu_j + I_{\{k \neq b(j)\}}\delta_{j,b(j)k} + \epsilon_{ijk} \tag{3}$$

$$\delta_j = \begin{pmatrix} \delta_{j,12} \\ \vdots \\ \delta_{j,1a_j} \end{pmatrix} \sim N_{a_j-1}\left(\begin{pmatrix} d_{t_{j1}t_{j2}} \\ \vdots \\ d_{t_{j1}t_{ja_j}} \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2/2 & \cdots & \tau^2/2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2/2 & \tau^2/2 & \vdots & \tau^2 \end{pmatrix} \right)$$

$$\epsilon_{ijk} \sim N(0, \sigma_j^2)$$

and for AD the model can be derived in similar way as in previous section

$$\bar{y}_{\cdot jk} = \mu_j + I_{\{k \neq b(j)\}}\delta_{j,b(j)k} + \bar{\epsilon}_{\cdot jk} \tag{4}$$

$$\delta_j = \begin{pmatrix}\delta_{j,12}\\ \vdots \\ \delta_{j,1a_j}\end{pmatrix} \sim N_{a_j-1}\left(\begin{pmatrix}d_{t_{j1}t_{j2}}\\ \vdots \\ d_{t_{j1}t_{ja_j}}\end{pmatrix}, \begin{pmatrix}\tau^2 & \tau^2/2 & \cdots & \tau^2/2\\ \vdots & \vdots & \ddots & \vdots \\ \tau^2/2 & \tau^2/2 & \vdots & \tau^2\end{pmatrix}\right)$$
$$\bar{\epsilon}_{\cdot jk} \sim N(0, \frac{\hat{\sigma}_j^2}{n_{jk}})$$

where y_{ijk} and $y_{\cdot jk}$ are the observed responses for IPD and AD as defined in previous section. The study specific baseline treatment is b and μ_j is the expected value for baseline treatment in study j. The parameter δ_{jbk} is the mean treatment effect for treatment k versus baseline treatment in study j, which are assumed to be realisations from normal distribution with mean d_{bk} and variance τ^2. Typically the first treatment in treatment list, so treatment A in our example, is selected to be reference treatment and treatment effects are estimated related to the reference treatment A and d_{Ak}s are considered as basic parameters and other parameters can be written as their functions(i.e. $d_{bk} = d_{Ak} - d_{Ab}$). The correlation between two random effects δ_{jk} is set to $\tau^2/2$ to induce equal between study variation for all treatment comparisons (i.e. Var(δ_{bk})=Var(δ_{Ak})=Var(δ_{Ab})).

3 The relation between two-way linear model and baseline contrast model

If we reparametrise the model 3 by substituting $\mu_j = \alpha_j + \theta_{b(j)} + u_{jb(j)}$ and $\delta_{j,b(j)k} = \theta_k + u_{jk} - (\theta_{jb(j)} + u_{jb(j)})$ we get

$$\begin{aligned} y_{ijk} &= \alpha_j + \theta_{b(j)} + u_{jb(j)} + I_{\{k\neq b(j)\}}(\theta_k + u_{jk} - (\theta_{jb(j)} + u_{jb(j)})) + \epsilon_{ijk} \\ &= \alpha_j + \theta_{b(j)} + u_{jb(j)} + I_{\{k\neq b(j)\}}(\theta_k - \theta_{jb(j)} + \tilde{u}_{jk}) + \epsilon_{ijk} \end{aligned}$$
$$\tilde{u}_{jk} = \begin{pmatrix}u_{j2} - u_{jb(j)}\\ \vdots \\ u_{ja_j} - u_{jb(j)}\end{pmatrix} \sim N_{a_j-1}\left(\begin{pmatrix}0\\ \vdots \\ 0\end{pmatrix}, \begin{pmatrix}2\sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2\\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \vdots & 2\sigma_u^2\end{pmatrix}\right) \quad (5)$$
$$\epsilon_{ijk} \sim N(0, \sigma_j^2)$$

From 5 we can see that essentially the only difference between models 1 and 3 is the assumption about the distribution of random effects.

3 Model fitting

The two models presented in previous section can be fitted in both frequentist and Bayesian framework. When frequentist based approaches are used for the models 1 or 3 several methods exists for estimating the variance components. In this paper we concentrate on maximum likelihood(ML) and restricted maximum likelihood(REML) methods. The maximum likelihood method is known to be biased when estimating the variance components. When REML is used, both models give equivalent results which is not the case with ML estimation. The baseline contrast model(3) is often fitted in bayesian framework using vague priors for the fixed effects and between study variance. As REML is the preferred method for estimating the variance components we will concentrate on the model 1 in the following section when discussing models with interactions.

3 Models with treatment-by-covariate interactions

In this sections we present models with treatment-by-covariate interactions. We are concentrating on continuous covariates, for example age, weight, blood pressure, or baseline value of the respose.

3 Model with treatment-by-covariate interaction

Let us first consider a model with treatment-by-covariate interaction for individual patient data:

$$y_{ijk} = \alpha_j + \theta_k + \gamma_j x_{ijk} + \beta_k x_{ijk} + u_{jk} + \epsilon_{ijk} \tag{6}$$

In this model γ_j is the fixed slope for study j and β_k is the fixed slope for treatment k, all the other components are equal to model 1. This model makes an assumtion that the slopes may vary across studies but the difference in slopes is fixed across treatments. In practice this could mean that in some trials the covariate has a higher effect on on the outcome for all treatments, but the effect between Again, like in model 1 we need a parameter condition for some of the fixed parameters. As we are more interested in the differences in covariate effects between treatments a natural choice is to set β_K to zero. Now the parameter γ_j is the slope for treatment K in study j and β_1 to β_{K-1} represent the difference in slopes compared to treatment K.

Let us now consider model 6 for aggregated data:

$$\begin{aligned} y_{\cdot jk} = \frac{\sum_{i=1}^{n_{jk}} y_{ijk}}{n_{jk}} &= \frac{\sum_{i=1}^{n_{jk}} (\alpha_j + \theta_k + \gamma_j x_{ijk} + \beta_k x_{ijk} + u_{jk} + \epsilon_{ijk})}{n_{jk}} \\ &= \frac{n_{jk}\alpha_j + n_{jk}\theta_k + \gamma_j \sum_{i=1}^{n_{jk}} x_{ijk} + \beta_k \sum_{i=1}^{n_{jk}} x_{ijk} + n_{jk}u_{jk} + \sum_{i=1}^{n_{jk}} \epsilon_{ijk}}{n_{jk}} \\ &= \alpha_j + \theta_k + \gamma_j \bar{x}_{\cdot jk} + \beta_k \bar{x}_{\cdot jk} + u_{jk} + \epsilon_{\cdot jk} \end{aligned} \tag{7}$$

If studies are reporting only mean covariate values per study , we observe only $\bar{x}_{\cdot j \cdot}$. If, in addition, we assume that due to randomization $x_{\cdot jk} \approx x_{\cdot j \cdot}$ for all k, model 7 becomes::

$$\begin{aligned} y_{\cdot jk} &= \alpha_j + \theta_k + \gamma_j x_{\cdot j \cdot} + \beta_k x_{\cdot j \cdot} + u_{jk} + \epsilon_{\cdot jk} \\ &= \lambda_j + \theta_k + \beta_k x_{\cdot j \cdot} + u_{jk} + \epsilon_{\cdot jk} \end{aligned}$$

As we observe only one covariate value per study the terms α_j and $\gamma_j x_{\cdot j\cdot}$ cannot be estimated separately. Therefore we write $\alpha_j + \gamma_j x_{\cdot j\cdot} = \lambda_j$. This model makes an additional assumption that the mean covariate values are similar for all treatments within study. Since we are working with randomized clinical trials this assumption should usually hold in practice.

3 Estimation of the interaction term

[12] In this section we will discuss the estimation of the interaction parameters β_k's and their variance. It is obvious that we lose something when the treatment-covariate interaction is estimated from the AD and next we will provide an exemplary calculation to assess the quantitative loss of estimating β_ks from AD instead of IPD. To simplify the algebra we will make some assumtions:
1) We will consider a fixed effect model($\sigma_u^2 = 0$)
2) We assume γ_j to be same for all studies
3) We assume that all studies have same number of treatments and all treatments have same number of patients
4) We assume that the residual variance $\sigma_j^2 = \sigma_\epsilon^2$ is the same for each study and
5) We assume that the covariate distribution within study is the same for all treatments.

The difference in slopes for treatment 1 and 2 is estimated by $\hat{\beta}_1 - \hat{\beta}_2$ and its variance for individual patient data is given by

$$\begin{aligned}\mathrm{var}_{\mathrm{IPD}}(\hat{\beta}_1 - \hat{\beta}_2) &= \mathrm{var}_{\mathrm{IPD}}(\hat{\beta}_1) + \mathrm{var}_{\mathrm{IPD}}(\hat{\beta}_2)\\ &= \frac{\sigma_\epsilon^2}{\sum_{j=1}^{J}\sum_{i=1}^{n_{j1}}(x_{ij1} - x_{..1})^2} + \frac{\sigma_\epsilon^2}{\sum_{j=1}^{J}\sum_{i=1}^{n_{j1}}(x_{ij1} - x_{..2})^2} \\ &= \frac{\sigma_\epsilon^2}{Jn\sigma_x^2} + \frac{\sigma_\epsilon^2}{Jn\sigma_x^2} = \frac{2\sigma_\epsilon^2}{Jn\sigma_x^2}\end{aligned} \tag{8}$$

where

$$\sigma_x^2 \approx \frac{\sum_{i=1}^{n}(x_{ij1} - x_{..1})^2}{n} \approx \frac{\sum_{i=1}^{n}(x_{ij1} - x_{..2})^2}{n}. \tag{9}$$

For aggregated data the variance of $\hat{\beta}_1 - \hat{\beta}_2$ is

$$\begin{aligned}\mathrm{var}_{\mathrm{AD}}(\hat{\beta}_1 - \hat{\beta}_2) &= \mathrm{var}_{\mathrm{AD}}(\hat{\beta}_1) + \mathrm{var}_{\mathrm{AD}}(\hat{\beta}_2)\\ &= \frac{\sigma_\epsilon^2/n}{\sum_{j=1}^{J}(x_{.j1} - x_{..1})^2} + \frac{\sigma_\epsilon^2/n}{\sum_{j=1}^{J}(x_{.j1} - x_{..2})^2}\\ &= \frac{\sigma_\epsilon^2}{Jn\sigma_{\bar{x}}^2} + \frac{\sigma_\epsilon^2}{Jn\sigma_{\bar{x}}^2} = \frac{2\sigma_\epsilon^2}{Jn\sigma_{\bar{x}}^2}\end{aligned}$$

where

$$\sigma^2_{\bar{x}} \approx \frac{\sum_{j=1}^{J}(x_{\cdot j1} - x_{\cdot\cdot 1})^2}{J} \approx \frac{\sum_{j=1}^{J}(x_{\cdot j2} - x_{\cdot\cdot 2})^2}{J}. \tag{10}$$

As the variance of x can be written as sum of between study variance of mean of x and within study variance of x we have $\sigma^2_x = \sigma^2_{\bar{x}} + \sigma^2_{x|\bar{x}}$. By substituting σ^2_x with $\sigma^2_{\bar{x}} + \sigma^2_{x|\bar{x}}$ in 8 we get

$$\begin{aligned} \text{var}_{\text{AD}}(\hat{\beta}_1 - \hat{\beta}_2) &= \frac{\sigma^2_{\bar{x}} + \sigma^2_{x|\bar{x}}}{\sigma^2_{\bar{x}}} \text{var}_{\text{IPD}}(\hat{\beta}_1 - \hat{\beta}_2) \\ &= \left(1 + \frac{\sigma^2_{x|\bar{x}}}{\sigma^2_{\bar{x}}}\right) \text{var}_{\text{IPD}}(\hat{\beta}_1 - \hat{\beta}_2) \end{aligned} \tag{11}$$

From 11 it can easily be seen that when within study variance of x is zero(i.e. all subjects within the study have the same covariate value) the variance of $\hat{\beta}_1 - \hat{\beta}_2$ estimated from aggregated data is the same as the variance estimated from the individual patient data. When variance of $\bar{x}$ approaches zero(i.e. all studies have the same covariate distribution and sample size is increasing) the variance estimated from aggregated data approaches infinity. Obviously, the interaction term can not be estimated from AD in this situation.

When we observe individual patient data for studies including treatment 1 and aggregated data for studies including treatment 2 the variance of $\hat{\beta}_1 - \hat{\beta}_2$ is simply

$$\begin{aligned} \text{var}_{\text{combined}}(\hat{\beta}_1 - \hat{\beta}_2) &= \text{var}_{\text{IPD}}(\hat{\beta}_1) + \text{var}_{\text{AD}}(\hat{\beta}_2) \\ &= \frac{\sigma^2_\epsilon}{Jn\sigma^2_x} + \frac{\sigma^2_\epsilon}{Jn\sigma^2_{\bar{x}}} \end{aligned}$$

3 Testing for ecological bias

In the AD studies we have only the aggregated values of the covariate and performing meta-regression with AD is subject to ecological bias. Ecological bias means that the effect of the covariate to the treatment effect is different in aggregated level than individual patient level. If we want to test wether ecological exists we can separate the regression coefficients for IPD and AD as proposed by Riley et. all in [9] The network meta-analysis model separating patient-level and study-level interactions can be written as

$$y^*_{ijk} = \alpha_j + \theta_k + \gamma_j x^*_{ijk} + \beta^w_k(x^*_{ijk} - \bar{x}_j) + \beta^a_k \bar{x}_j + u_{jk} + \epsilon^*_{ijk} \tag{12}$$

where y^*_{ijk}, ϵ^*_{ijk} and x^*_{ijk} are y_{ijk}, ϵ_{ijk}, and x_{ijk} for IPD studies and $y_{\cdot jk}$, $\epsilon_{\cdot jk}$, and $x_{\cdot j \cdot}$ for AD studies, β^w_k is the within-study regression coefficient, β^a_k is the accross-study regression coefficient, and $\bar{x}_j$ is the mean covariate value in study j. Note that for AD studies $\bar{x}_j = x_{ijk}$ following $\bar{x}_j - x_{ijk} = 0$ and therefore the within-study regression coefficient cancels out for AD studies and only IPD studies are used to estimate the within-study regression coefficient. From model 12 the presens of ecological bias can be tested by hypothesis $H : \beta_k^a - \beta_k^w = 0$. If all studies provide IPD model 12 can be written as

$$\begin{aligned} y_{ijk} &= \alpha_j + \theta_k + \gamma_j x_{ijk} + \beta_k^w x_{ijk} + (\beta_k^a - \beta_k^w)\bar{x}_j + u_{jk} + \epsilon_{ijk} \\ &= \alpha_j + \theta_k + \gamma_j x_{ijk} + \beta_k^w x_{ijk} + \beta_k^{a-w}\bar{x}_j + u_{jk} + \epsilon_{ijk} \end{aligned} \tag{13}$$

This is the same model as model 6 with an additional term $\beta_k^{a-w}\bar{x}_j$ which is the effect of mean covariate value in study j. Therefore testing ecological bias is the same as testing the effect of mean covariate value in IPD. When ecological bias is absent $\beta_k^{a-w}\bar{x}_j = 0$ and model 13 is essentially the same as model 6.

3 Application to the example data

3 Implementation

All methods presented above were applied to the example data. For the frequentist models SAS proc mixed[10] was used with restricted maximum likehood estimation(REML). For bayesian analyses we used JAGS[8]. In bayesian analyses we chose uniform prior distribution for τ,$\tau \sim uniform(0, 2)$ and gamma distribution for the inverse of σ_{jk}^2. For all other parameters we used normal priors with mean zero and variance 10^5. We ran three chains with different starting values and with 100 000 iterations for each with 50 000 burn in samples for each. All methods were applied to 3 different types of data. First analysis (IPD) was all studies with IPD, second analysis (IPD-AD) was studies 1-14 with AD and studies 15-21 with IPD, and third analysis (AD) was all studies with AD.

3 Results

Table 2 shows the relative treatment effects compared to placebo from the analyses without treatment-by-covariate interactions. We present the results for models 1 and 3 using both REML and ML estimation and in addition we present the results for model 3 using bayesian methods. In model 3 analysis the random effect was added to the relative treatment effects as in model 1 we put random effects on each treatment arm within study, therefore the τ^2 is expected to be twice as large as σ_u^2. The point estimates for the treatment effects are in close agreement with all compared methods. From frequentist methods the REML estimation gives wider confidence intervals and larger between study variance σ_u^2 than ML estimation. The ML estimation for model 3 was more conservative than for model 1, giving wider confidence intervals and larger between study variance estimate. The results from Bayesian analysis are close to the REML estimation. With all models and estimation techniques the results from IPD, IPD-AD, and AD are in close agreement.

		Estimate and 95% Confidence interval		
Model	Parameter	IPD	IPD-AD	AD
Models 1 and 3	θ_A	0.689(0.525,0.852)	0.689(0.525,0.853)	0.69(0.526,0.854)
REML	θ_B	0.543(0.419,0.667)	0.544(0.42,0.668)	0.544(0.421,0.668)
	θ_C	0.592(0.294,0.891)	0.594(0.293,0.894)	0.594(0.294,0.893)
	θ_D	0.555(0.364,0.746)	0.556(0.364,0.748)	0.556(0.365,0.748)
	θ_E	0.236(-0.085,0.557)	0.235(-0.084,0.553)	0.235(-0.083,0.552)
	θ_F	0.448(0.289,0.607)	0.449(0.291,0.608)	0.45(0.291,0.608)
	$\theta_A - \theta_B$	0.145(-0.042,0.332)	0.145(-0.042,0.332)	0.146(-0.041,0.333)
	σ_u^2	0.0148	0.0148	0.0148
Model 1 ML	θ_A	0.678(0.555,0.8)	0.679(0.557,0.801)	0.681(0.559,0.803)
	θ_B	0.545(0.452,0.639)	0.546(0.453,0.639)	0.547(0.454,0.639)
	θ_C	0.571(0.346,0.796)	0.573(0.345,0.8)	0.573(0.345,0.8)
	θ_D	0.554(0.413,0.695)	0.555(0.414,0.696)	0.556(0.415,0.697)
	θ_E	0.244(-0.004,0.492)	0.243(-0.003,0.489)	0.243(-0.003,0.489)
	θ_F	0.441(0.327,0.556)	0.443(0.329,0.557)	0.444(0.33,0.558)
	$\theta_A - \theta_B$	0.132(-0.005,0.269)	0.132(-0.004,0.269)	0.134(-0.002,0.271)
	σ_u^2	0.005	0.004	0.005
Model 3 ML	d_{PA}	0.683(0.545,0.821)	0.684(0.546,0.822)	0.686(0.548,0.824)
	d_{PB}	0.544(0.439,0.649)	0.545(0.44,0.65)	0.545(0.441,0.65)
	d_{PC}	0.58(0.326,0.834)	0.582(0.326,0.838)	0.582(0.326,0.837)
	d_{PD}	0.555(0.395,0.715)	0.556(0.395,0.716)	0.556(0.396,0.716)
	d_{PE}	0.241(-0.035,0.516)	0.239(-0.034,0.512)	0.239(-0.034,0.512)
	d_{PF}	0.445(0.313,0.577)	0.447(0.315,0.578)	0.447(0.316,0.578)
	d_{BA}	0.139(-0.017,0.296)	0.139(-0.017,0.295)	0.141(-0.015,0.297)
	$\tau^2/2$	0.008	0.008	0.008
		Posterior median and 95% Credibility interval		
Model	Parameter	IPD	IPD-AD	AD
Model 3 Bayesian	d_{PA}	0.69(0.517,0.864)	0.692(0.516,0.869)	0.69(0.518,0.865)
	d_{PB}	0.545(0.413,0.676)	0.545(0.414,0.677)	0.544(0.413,0.675)
	d_{PC}	0.596(0.281,0.914)	0.595(0.283,0.915)	0.594(0.281,0.916)
	d_{PD}	0.555(0.349,0.758)	0.556(0.352,0.763)	0.556(0.352,0.76)
	d_{PE}	0.237(-0.104,0.571)	0.234(-0.101,0.567)	0.234(-0.1,0.569)
	d_{PF}	0.449(0.281,0.621)	0.451(0.28,0.622)	0.449(0.281,0.619)
	d_{BA}	-0.145(-0.349,0.054)	-0.147(-0.351,0.052)	-0.146(-0.348,0.052)
	$\tau^2/2$	0.0160	0.0163	0.0160

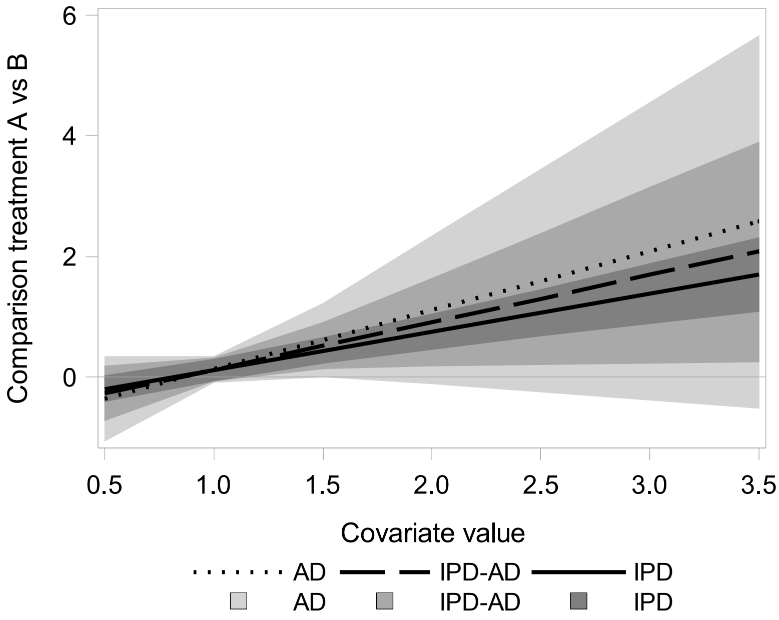
Table 2: Parameter estimates for models without treatment-covariate interactions

Table 3 shows the parameter estimates from the model with treatment effect modifying covariate x. Figure 2 shows the estimated treatment difference and 95% confidence interval between treatment A and B at different covariate values. With all three data types the estimates and confidence intervals are in close agreement when the covariate value is close to the mean covariate values. As expected from analytical consideration in section 3.2.2, if the treatment difference is estimated at lower or higher covariate values the IPD analysis has much narrower confidence intervals than the analyses containing AD. Also, the point estimates vary more at more extreme covariate values.

Model	Parameter	Estimate and 95% CI		
		IPD	IPD- AD	AD
Model 6	θ_A	0.658(0.501,0.815)	0.656(0.492,0.82)	0.662(0.476,0.849)
REML	θ_B	0.542(0.423,0.661)	0.541(0.412,0.671)	0.543(0.406,0.68)
	θ_C	0.62(0.319,0.922)	0.81(0.093,1.526)	0.807(0.077,1.538)
	θ_D	0.539(0.356,0.722)	0.539(0.347,0.731)	0.531(0.322,0.74)
	θ_E	0.243(-0.067,0.554)	0.062(-1.028,1.152)	0.063(-1.056,1.182)
	θ_F	0.431(0.279,0.583)	0.432(0.272,0.592)	0.441(0.265,0.618)
	$\theta_A - \theta_B$	0.116(-0.063,0.295)	0.115(-0.074,0.304)	0.12(-0.087,0.327)
	β_A	0.608(0.397,0.818)	0.66(0.412,0.908)	0.821(-0.262,1.905)
	β_B	-0.023(-0.176,0.129)	-0.124(-0.832,0.584)	-0.158(-0.925,0.609)
	β_C	-0.168(-0.554,0.217)	-1.064(-4.064,1.936)	-1.069(-4.155,2.017)
	β_D	0.253(0.01,0.496)	0.341(-0.193,0.874)	-0.012(-2.376,2.352)
	β_E	-0.121(-0.542,0.299)	3.096(-16.113,22.305)	3.073(-16.731,22.877)
	β_F	0.077(-0.109,0.263)	0.206(-0.144,0.556)	-0.003(-1.578,1.572)
	$\beta_A - \beta_B$	0.631(0.398,0.864)	0.784(0.045,1.524)	0.98(-0.274,2.233)
	σ_u^2	0.0129	0.0147	0.0168

Table 3: Parameter estimates for model with treatment-covariate interaction

Figure 2: Estimated treatment effect A vs B at different covariate values with 95%-confidence bands



4 Simulation

To compare the proposed models and estimation treatment effects and treatment-by-covariate interaction under a random effect model we performed a simulation study.

4 Simulation setting

To compare the models presented in section 3 we performed a simulation study for network meta-analysis with no additional covariates and with one effect modifying covariate. The data for the network meta analysis model were generated with 4 treatments and different number of studies. The number of IPD studies was 5 in all simulations and number of AD studies varied ($J_{AD} = 5, 10, 20, 30$). The IPD studies were generated with treatments A, B, and C and AD studies were generated with treatments B, C, and D. The within study standard deviation was 1 and between study standard deviation varied $\sigma_u = 1/8, 1/4, 1/2$, corrwsponding to moderate, substantial and large heterogeneity values. For each combination of J_{AD} and σ_u we generated 10000 individual patient datasets, which were aggregated for the network meta-analysis when applicable. For model with the treatment-effect modifying covariate the true value of $\beta_1 - \beta_2$ was 0.66 and var(x) was 1 with different combinations of var($\bar(x)$) and var($x|\bar{x}$). In the simulated scenarious 1 to 8 the corresponding values for var($\bar{x}$) were 0,0.05,0.1,0.2,0.4,0.6,0.8,1 and and for var($x|\bar{x}$)=1-var($\bar{x}$). We simulated the covariate values in two stages, first we simulated the mean covariate value $\tilde{x}$ from $N(0, \sigma^2_{\bar{x}})$ and second the individual patient covariate value from $N(\tilde{x}, \sigma^2_{x|\bar{x}})$. A new single study comparing A nd B would have 90% power to detect the interaction with the values used in simulation.

4 Results

4 No treatment effect modifying covariates

Figure 3 sows the estimated between study standard deviation(σ_u) for model 1 and 3. When REML is used for the estimation no bias can be seen in the estimation of σ_u. With ML estimation both models are biased downwards and model 1 is more biased than model 3. With model 1 no differences between the different data types can be seen when REML is used and ML estimation all datatypes are in close agreement except when between study standard deviation is small($\sigma_u = 1$).

Figure 4 shows the type 1 error rate under the null hypothesis. The desired level of type 1 error rate is 5%. Model 1 with REML holds the pre-specified α-level best, varying between 5.2% and 7.1%. When ML estimation is used, both models are too liberal, for model 1 the type 1 error rate varies between 12.5% and 17.5% and for model 3 between 5.9% and 12.7%. All models are less liberal when the number of studies is increasing. This is due to the more accurate estimation of between study standard deviation and the use of normal approximation for the hypothesis testing.

4 One treatment effect modifying covariates

Figure 5 shows the power for detecting the treatment-by-covariate interaction with all 8 different scenarios covariate distributions. The grey reference line represents the power of detecting the treatment-by-covariate interaction in a single study directly comparing treatments A and B. If all studies provide IPD

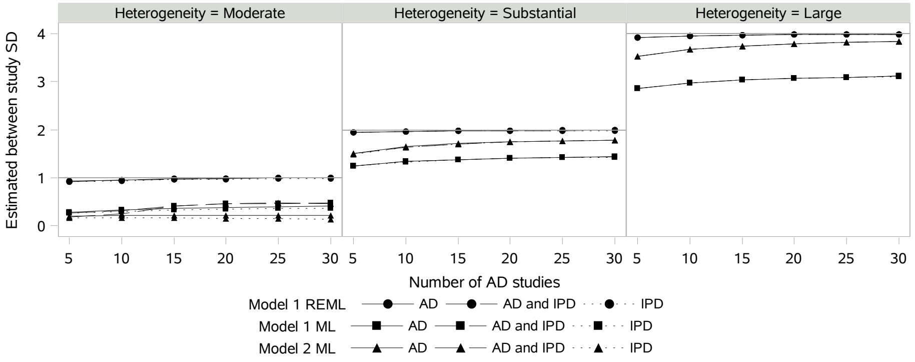


Figure 3: Estimated between study variance in two-way-linear-mixed model and baseline contrast model

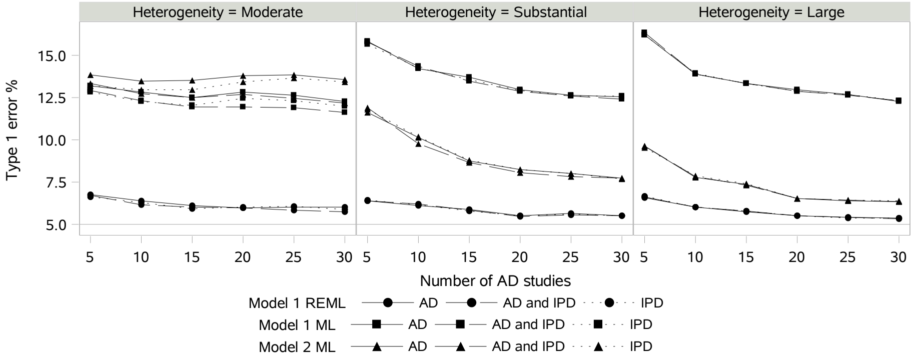


Figure 4: Type 1 error rate for the treatment difference A vs B in two-way-linear-mixed model and baseline contrast model

the power for detecting interaction is always 100% for first six studies, which is natural as having IPD for all studies the sample size of is at least ten times as large as with one single study. When some or all of the studies provide only AD the power to detect the interaction is close to 5% in scenario 1(the true mean of covariate distribution is same in all studies), which is the type 1 error rate for the test. As the variation in mean covariate values increases the power of analysis with IPD-AD or AD alone increases, and when there is no within study variation in covariate values (scenario 8) the all three methods give same results as expected from analytical consideration in Section 3.2.2.

In scenario 3 in the simulation the covariate value variances are on the same level as in the example data. From the simulation results we can see that in order to have 90% power for the treatment-by-covariate test with IPD-AD, we need 20 AD studies if the other heterogeneity is moderate.

5 Discussion

In this article, we provided a framework for the network meta-analysis of clinical trials combining individual patient data and aggregated data. The proposed

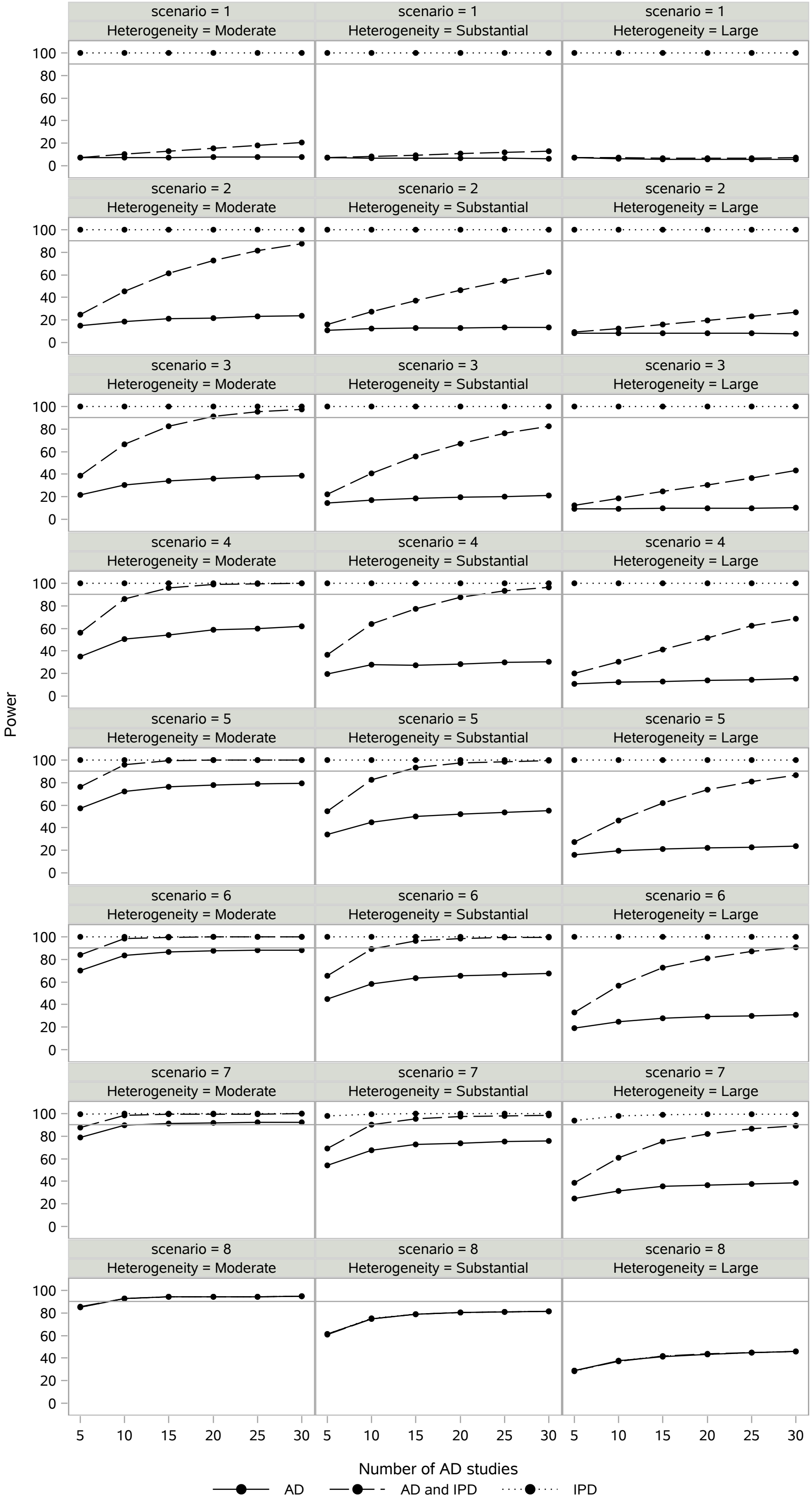


Figure 5: Power to detect the interaction with equal between- and within-study covariate variances, $\bar{x}$=4 and var($x|\bar{x}$)=4

approach is based on the ANOVA framework. Same type of models have been proposed in the model framework based on baseline contrast mode[3, 2]. In our opinion the parametrization of the ANOVA framework is easier to understand and implement than the baseline contrast parametrisation and therefore we prefer the ANOVA approach. When adding covariates and treatment-by-covariate interactions to the models it is straightforward with the ANOVA framework whereas with the baseline contrast model the additional complication of the model parametrisation inherits to the interaction terms as well even though the treatment-covariate interactions are typically fitted as fixed effects.

We focused on situation where one of the treatments in the network has individual patient data and we are comparing with a treatment providing only aggregated data. This kind of situation can arise in in pharmaceutical industry, when two compounds are developed by different companies and it hasn't been possible to use the competitor compound as a comparator in the in-house studies.

We showed by simulation that when using the REML estimation for continuous outcome the ANOVA model holds the pre-specified type 1 error rate and the estimated between-study variance is closer to the true value than when using maximum likelihood estimation with the ANOVA model or with baseline contrast model. It was also shown by simulation that when no treatment-by-covariate interactions are present in the model the aggregated data is sufficient for performing the network meta-analysis.

We provided the analytical calculations for the variance of the interaction for fixed effect model and we investigated the situation under random effect model by simulation. When network meta-analysis is performed with aggregated data and additional treatment effect modifying covariates are included in the model the distribution of the covariate values between studies is crucial. If all studies have the same covariate distribution most of the information about the interaction will be lost when when the data is aggregated.

We systematically evaluated the number of AD studies needed to estimate the treatment-by-covariate with certain level of power. Our simulation study shows that both the covariate heterogeneity and other heterogeneity in the network have an effect on the power. The more heterogeneinty in the mean covariate values the higher the power and the more other heterogeneity the lower the power. In other words this means that in order to use AD for analysing the treatment-by-covariate interactions, the studies should be as heterogeneous as possible when it comes to the covariate of interest but as homogeneous as possible respect to all other covariates.

We only covered continuous and normally distributied outcomes in our paper. When other type of responses are modelled, both the estimation of between study variance and treatment covariate interactions is affected. First, we can't use REML for other than normal data and the estimation of between study variation with will be biased. Second when the link funktion is not identity function the derivation of AD model is not as straightforward as with linear models.

In summary, we conclude that when treatment-by-covariate interactions are of interest one should always try to get access to some IPD for all treatment arms. If the NMA is limited to the use of AD the analyst should carefully evaluate the covariate distributions of the AD studies before making any statemants on the interactions.

References

[1] I Demin, B Hamrén, O Luttringer, G Pillai, and T Jung. Longitudinal model-based meta-analysis in rheumatoid arthritis: an application toward model-based drug development. *Clinical Pharmacology & Therapeutics*, 92(3):352–359, 2012.

[2] 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):914–930, 2013.

[3] Jeroen P Jansen. Network meta-analysis of individual and aggregate level data. *Research Synthesis Methods*, 3(2):177–190, 2012.

[4] Byron Jones, James Roger, Peter W Lane, Andy Lawton, Chrissie Fletcher, Joseph C Cappelleri, Helen Tate, and Patrick Moneuse. Statistical approaches for conducting network meta-analysis in drug development. *Pharmaceutical statistics*, 10(6):523–531, 2011.

[5] Paul C Lambert, Alex J Sutton, Keith R Abrams, and David R Jones. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of clinical epidemiology*, 55(1):86–94, 2002.

[6] Guobing Lu and AE Ades. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association*, 101(474), 2006.

[7] H. P. Piepho, E. R. Williams, and L. V. Madden. The use of two-way linear mixed models in multitreatment meta-analysis. *Biometrics*, 68(4):1269–1277, 2012.

[8] Martyn Plummer. rjags: Bayesian graphical models using mcmc. *R package version*, 2(0), 2011.

[9] Richard D. Riley, Paul C. Lambert, Jan A. Staessen, Jiguang Wang, Francois Gueyffier, Lutgarde Thijs, and Florent Boutitie. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine*, 27(11):1870–1893, 2008.

[10] Inc SAS Institute. Sas® 9.3, 2011.

[11] Stephen Senn. The many modes of meta. *Drug Information Journal*, 34(2):535–549, 2000.

[12] MC Simmonds and JPT Higgins. Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. *Statistics in medicine*, 26(15):2982–2999, 2007.

[13] Lesley A. Stewart. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine*, 14(19):2057–2079, 1995.

[14] Alex J Sutton, Denise Kendrick, and Carol AC Coupland. Meta-analysis of individual-and aggregate-level data. *Statistics in medicine*, 27(5):651–669, 2008.

3 Estimating relative efficacy in acute postoperative pain: network meta-analysis is consistent with indirect comparison to placebo alone

Abstract:

In this manuscript a network meta-analysis of pain medication was conducted and compared with Cochrane indirect analyses. The network analysis was based on clinical trials of acute postoperative pain. The 261 trials published between 1966 and 2016 included 39,753 patients examining 52 active drug and dose combinations (27,726 given active drug and 12,027 placebo), in any type of surgery (72% dental). The outcome of interest was a binomial endpoint measuring if certain degree of pain relief was achieved or not. The network meta-analsis was conducted using generalized mixed models with arm-based modelling.

This chapter is published as:

Moore, R.A., Derry, S., Wiffen, P.J., Banerjee, S., Karan, R., Glimm, E., Wiksten, A., Aldington, D., Eccleston, C., 2018. Estimating relative efficacy in acute postoperative pain: network meta-analysis is consistent with indirect comparison to placebo alone. PAIN 159, 2234–2244. https://doi.org/10.1097/j.pain.0000000000001322

4 Nonproportional Hazards in Network Meta-Analysis: Efficient Strategies for Model Building and Analysis

ScienceDirect

Contents lists available at **sciencedirect.com**
Journal homepage: **www.elsevier.com/locate/jval**

Methodology

Nonproportional Hazards in Network Meta-Analysis: Efficient Strategies for Model Building and Analysis

Anna Wiksten, MSc, Neil Hawkins, PhD, Hans-Peter Piepho, PhD, Sandro Gsteiger, PhD*

ABSTRACT

Objectives: To develop efficient approaches for fitting network meta-analysis (NMA) models with time-varying hazard ratios (such as fractional polynomials and piecewise constant models) to allow practitioners to investigate a broad range of models rapidly and to achieve a more robust and comprehensive model selection strategy.

Methods: We reformulated the fractional polynomial and piecewise constant NMA models using analysis of variance–like parameterization. With this approach, both models are expressed as generalized linear models (GLMs) with time-varying covariates. Such models can be fitted efficiently with standard frequentist techniques. We applied our approach to the example data from the study by Jansen et al, in which fractional polynomial NMA models were introduced.

Results: Fitting frequentist fixed-effect NMAs for a large initial set of candidate models took less than 1 second with standard GLM routines. This allowed for model selection from a large range of hazard ratio structures by comparing a set of criteria including Akaike information criterion/Bayesian information criterion, visual inspection of goodness-of-fit, and long-term extrapolations. The "best" models were then refitted in a Bayesian framework. Estimates agreed very closely.

Conclusions: NMA models with time-varying hazard ratios can be explored efficiently with a stepwise approach. A frequentist fixed-effect framework enables rapid exploration of different models. The best model can then be assessed further in a Bayesian framework to capture and propagate uncertainty for decision-making.

Keywords: fractional polynomial models, network meta-analysis, nonproportional hazards time-to-event data, piecewise exponential models.

VALUE HEALTH. 2020; 23(7):918–927

Introduction

Time-to-event (TTE) data showing nonproportional hazards are becoming increasingly common in network meta-analysis. Two major reasons are the inclusion of trials with relatively long follow-up and the advent of interventions with novel mechanisms of action. For example, the proportionality between hazard curves for overall survival or progression-free survival may be a reasonable approximation for short durations, but the assumption can become unrealistic as follow-up time increases.[1,2] In other instances, an innovative new treatment class can make the proportional hazards (PH) assumption unlikely. Cancer immunotherapy, for example, typically shows a delayed separation of survival curves and a long-term survival benefit when compared with chemotherapy.[3] PH cannot hold in such situations. Also, it is worth noting that composite endpoints such as progression-free survival are more likely to display nonproportional hazard ratios (HRs) than their individual component parts may suggest.[4]

Network meta-analysis (NMA) requires particular care regarding the modeling of hazard ratios over time for 2 main reasons. First, NMA models may combine trials with differences in follow-up time. Within a single randomized controlled trial, the hazard ratio can be interpreted as a weighted average of hazard ratios over small time intervals (on the log scale).[5] Therefore, the overall hazard ratio remains a valid (although arguably limited) summary and is reported as a primary measure of treatment effect even if hazards are not proportional throughout the trial period (see, for example, recent cancer immunotherapy trials).[6-8] However, the overall hazard ratio depends on the follow-up time of the trial. Therefore, the synthesis of hazard ratios from different trials violating the PH assumption could be confounded. Second, NMA results often feed into health economic models involving long-term extrapolations. Modeling of treatment effects beyond the observed data represents a major challenge whether PH holds or not. But time-varying hazard ratios can heavily affect conclusions and need careful consideration.

* Address correspondence to: Sandro Gsteiger, PhD, F. Hoffmann-La Roche Ltd, Division Pharma, Global Access, Grenzacherstrasse, CH-4070 Basel, Switzerland.
Email: sandro.gsteiger@roche.com

NMAs with TTE data are most commonly based on hazard ratio estimates.[9] Two simple alternative approaches that substitute different consistency assumptions are area under the curve (AUC) models and accelerated failure time (AFT) models. For AUC models, treatment effects are summarized by mean differences in the area under the survival curve over a prespecified interval.[10] Such models are very general because they do not require any parametric assumptions (unless follow-up times are very different, in which case some tail area estimation using parametric models may be needed). The estimates per se are meaningful because the AUC estimates the average restricted mean survival time. However, the approach does not allow for longer-term extrapolation, and the synthesis imposes the constraint that differences in mean survival are consistent across trials and independent of differences in absolute survival. Another alternative replaces the PH assumption by a proportionality assumption on the time axis itself (ie, the treatment effects are modeled via acceleration factors in an AFT model).[11] In this case, the synthesis imposes the constraint that ratios of acceleration factors are consistent across trials and independent of differences in absolute survival. This may work in some cases; however, for example, the pattern seen with cancer immunotherapy versus chemotherapy may not be accurately described with AFT models.

More general approaches model the hazard ratio curves over time. A prominent approach introduced by Jansen[12] uses fractional polynomials to describe the log-hazards. This also implies that the log-hazard ratios will be fractional polynomials of the same order. In this case, the synthesis imposes the constraint that all (two in a first-order model, three in a second-order model, etc) of the relevant parameters in the polynomial are independently consistent across trials and independent of differences in absolute survival. The approach is very flexible, but fractional polynomials can be difficult to use in practice (sensitivity to starting values, potential convergence issues). In addition, modeling the observed survival curves more precisely will not guarantee valid extrapolations; more complex models may lead to biologically implausible long-term predictions. Under such circumstances, either simpler models may be preferred or additional assumptions such as those described by Jackson et al[13] may be applied to ensure plausible predictions. Piecewise constant models provide another option: the PH assumption is made within segments, but not overall.[14] Typically, baseline hazards are modeled with the exponential distribution. Although the approach is quite simple, any desired level of flexibility can be achieved by increasing the number of segments. More recently, Freeman and Carpenter[2] presented an NMA using the Royston-Parmar models (restricted cubic spline models). The increased flexibility comes at the price of considerable additional complexity.

So far, the fractional polynomial model has been implemented using Bayesian Markov chain Monte Carlo techniques. Although clearly the Bayesian framework provides many appealing properties, complex models may be time-consuming to fit using Markov chain Monte Carlo simulation methods, and model building may not be straightforward. To explore many candidate models quickly, efficient strategies for model building and selection are needed for complex TTE NMAs. In this step, a frequentist framework providing point and interval estimates is often sufficient.

In this work, we reformulate NMA models with time-varying hazard ratios using analysis of variance (ANOVA) parameterization.[15] This approach allows fitting the corresponding fixed-effect NMAs in a frequentist framework via standard generalized linear model (GLM) routines, which are available in all major statistical packages such as R and Stata.[16,17] We show how to formulate fractional polynomials and piecewise exponential (PWE) models in this framework and how PH models can be expressed as special cases. The efficiency of the available GLM fitting routines allows exploring many such complex TTE NMA models very quickly. We illustrate the approach by reanalyzing the non–small cell lung cancer (NSCLC) example from the original article by Jansen,[12] in which fractional polynomial NMAs were introduced. Using the same data as Jansen will ease comparison of approaches, but it should be noted that the evidence base in NSCLC has changed considerably since.[18]

Methods

GLMs for Grouped Survival Data

From published Kaplan-Meier curves, the so-called Guyot algorithm allows the approximation of the underlying individual participant data (“pseudo” IPD).[19] For model fitting, it is convenient to group this underlying individual participant data into intervals. For study j and treatment k, we thus obtain grouped survival data given by the number of events r_{jkt}, and the number at risk n_{jkt} in a time interval $[t - \Delta t, t]$.

Prentice and Gloeckler showed that grouped survival data can be modeled with a binomial likelihood $r_{jkt} \sim Bin(p_{jkt}, n_{jkt})$ and complementary log-log link function,

$$\text{cloglog}\left(p_{jkt}\right) = \eta_{jkt} + \ln\,(\Delta t_{jkt}),$$

where η_{jkt} is the linear predictor for treatment k in study j at time t and $\ln(\Delta t_{jkt})$ is the offset term accounting for different lengths of time intervals.[20] The hazard function h_{jkt} of an underlying continuous-time model (with survivor function $S(t)$) relates to the event probability p_{jkt} via

$$p_{jkt} = \frac{S(t-\Delta t) - S(t)}{S(t-\Delta t)} = 1 - e^{-\int_{t-\Delta t}^{t} h(u)du} \cong 1 - e^{-\Delta t \cdot h_{jkt}}.$$

The approximation in the last step above assumes the hazard is constant over the interval $[t - \Delta t, t]$, which should be acceptable if the time steps are relatively small. Transforming this expression leads to the following approximation (corresponding to equation [8] from Jansen[12]),

$$h_{jkt} \cong -\ln\left(1 - p_{jkt}\right) \Big/ \Delta t_{jkt},$$

which shows that

$$\ln\left(h_{jkt}\right) \cong \text{cloglog}\left(p_{jkt}\right) - \ln\,(\Delta t_{jkt}) = \eta_{jkt}\cdot$$

This shows that any survival model with linear log-hazard function can be fitted as a GLM for grouped survival data. Note that “linear” means the unknown model parameters enter the log-hazard function linearly; it does not refer to the time dependency, which can be of any shape.

Specifying Time-Varying Hazard Ratio NMA Models Using Arm-Based Parameterization

The GLM approach for grouped survival data allows for fitting a large range of NMA models with time-varying hazard ratios. This is best seen when using the arm-based or ANOVA

Table 1. Models applied to the illustrative example and AIC values from frequentist fixed-effect model fits.

Model	$g_0(t)$	$g_1(t)$	$g_2(t)$	AIC
Exponential (PH)	1			1053.9
PH with Weibull baseline survival ($\theta_1 = 0$)	1	$\ln(t)$		958.3
First-order FP, $p_1 = 1$, Gompertz	1	t		969.1
First-order FP, $p_1 = 0$, Weibull	1	$\ln(t)$		955.1
First-order FP, $p_1 = -2$	1	t^{-2}		924.5
Second-order FP, $p_1 = -2$, $p_2 = 1$	1	t^{-2}	t	866.5
Piecewise exponential with 1 cut point	1	$I(t > 2)$		920.6
Piecewise exponential with 2 cut points	1	$I(2 < t \le 10)$	$I(t > 10)$	895.0
Piecewise exponential with 2 cut points	1	$I(2 < t \le 12)$	$I(t > 12)$	853.4

AIC indicates Akaike information criterion; FP, fractional polynomial; PH, proportional hazards.

parameterization of NMA.[21,22] The general form for η_{jkt} in a fixed-effect NMA with time-varying hazard ratios is

$$\eta_{jkt} = \sum_{m=0}^{M} (\alpha_{mj} + \theta_{mk}) g_m(t),$$

where α_{mj} are the study-specific coefficients for study j, the θ_{mk} are the treatment-specific coefficients for treatment k, M is the number of time-varying terms, and $g_m(t)$ are a set of time-varying "basis" functions. For example, by setting $g_m(t) = t^{p_m}$ for a set of prespecified exponents p_m, we obtain the fractional polynomial (of Mth order) NMA models introduced in Jansen.[12] Models with as few as 1 or 2 time-varying terms lead to great flexibility of the resulting hazard ratio shapes, such as monotonic increasing/decreasing, bathtub, and inverse-bathtub shaped. For practical purposes, M = 1, 2 will often be sufficient.

Similarly, we can express PWE models via step functions. For example,

$$\eta_{jkt} = \alpha_{0j} + \theta_{0k} + (\alpha_{1j} + \theta_{1k}) I(t \ge C_1)$$

provides a model with 1 cut point at $t = C_1$. In this equation, α_{0j} corresponds to the log-hazard over the first segment and α_{1j} to the difference in log-hazards between the second and the first segments for study j for the reference treatment (whether included in the study or not). The term θ_{0k} corresponds to the contrast between treatment k and the reference treatment (regardless of study) over the first segment, and θ_{1k} is the difference in contrasts between the second and first segment. PH models correspond to the special case

$$\eta_{jkt} = \sum_{m=0}^{M} \alpha_{mj} g_m(t) + \theta_{0k},$$

where the higher-order θ_{mk} terms are restricted to zero. Also, the Gompertz, Weibull, and exponential models fit into this framework (Table 1).

The vectors (θ_{0k}, θ_{1k}, ..., θ_{Mk}) model the M+1 dimensional treatment effect for comparing treatment k to a reference treatment (k = 1) and θ_{m1} = 0 for all m for identifiability. The hazard ratio comparing treatment B versus treatment A at time t in this NMA model is

$$HR_{AB}(t) = e^{\sum_{m=0}^{M} (\theta_{mB} - \theta_{mA}) g_m(t)}.$$

Imposing consistency at the parameter level (θ_{mk}) leads to consistency of log-hazard ratios for all time points t, as shown in Ouwens et al.[23]

When using fixed-effect models, the more common contrast-based parametrization of NMA and the arm-based NMA parameterization will be equivalent. The contrast-based formulation can be obtained by reparameterization, namely,

$$\eta_{jkt} = \begin{cases} \sum_{m=0}^{M} \mu_{mjb} g_m(t), \text{ if } k = b, b = A, B, C, \ldots \\ \sum_{m=0}^{M} (\mu_{mjb} + d_{mbk}) g_m(t), \text{if } k \text{ alphabetically after } b \end{cases}$$

where $\mu_{mjb} = \alpha_{mj} + \theta_{mb}$ and $d_{mbk} = \theta_{mk} - \theta_{mb}$. In this notation, b stands for the baseline treatment in study j (and treatments have been suitably ordered and labelled A, B, C,...). The equivalence between the 2 parameterizations has been discussed in more detail for models without time-varying terms,[15,21] and these results extend naturally to our case.

It should be noted that the fixed study effects in our model ensure that all inference is based on within-study information (contrasts). This means the approach does respect randomization, which is an important feature of valid cross-trial synthesis methods. Other authors use a different notion of "arm-based models." For them, arm-based models are formulated in a way that recovers arm-level interstudy information, for example, by putting an exchangeability assumption on arm-level (absolute) outcomes from common treatments.[24-26]

Extension to Random-Effects Models

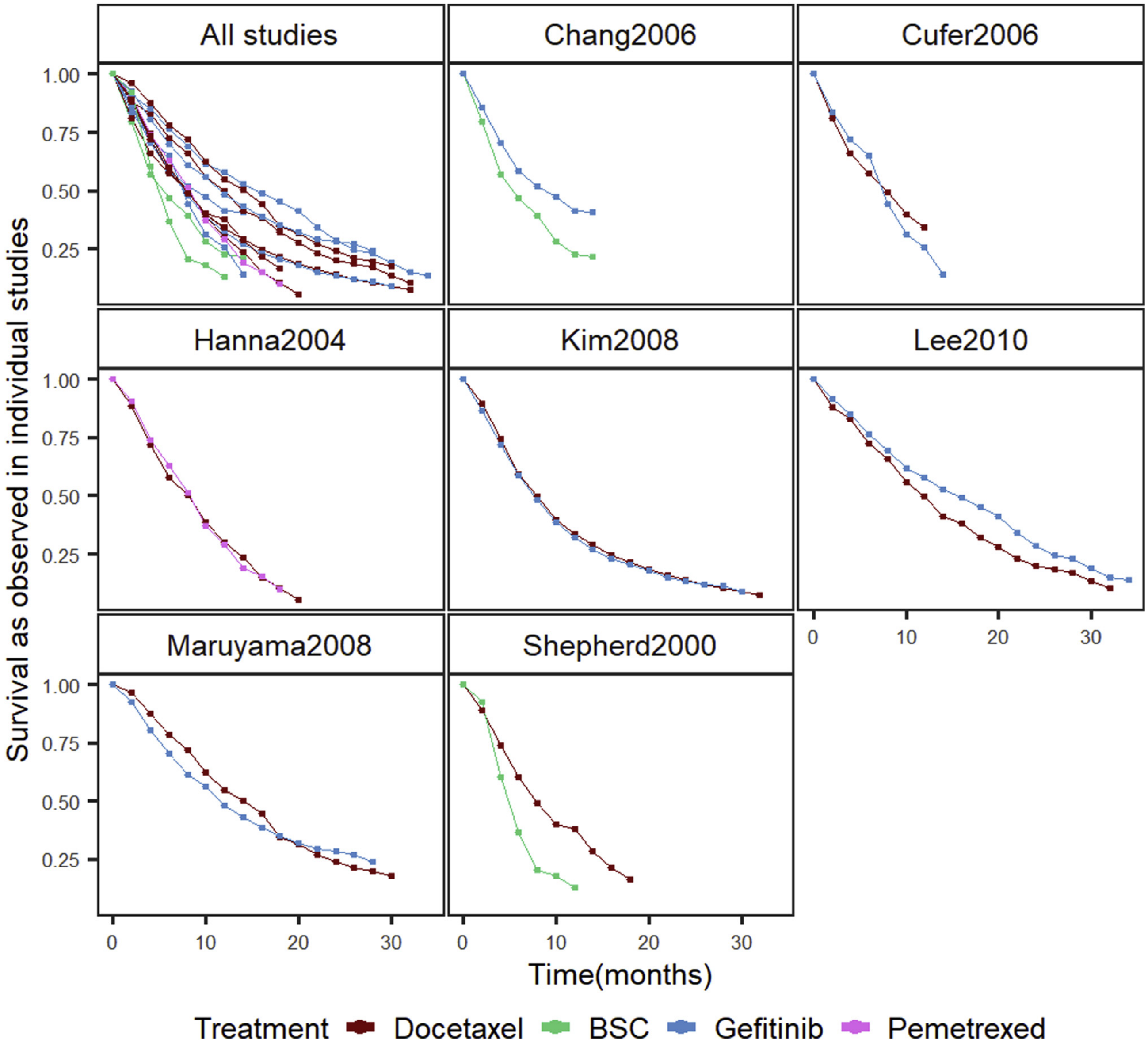
The model can be extended to account for between-trial heterogeneity by adding random treatment by study interaction terms.[15,21,27] In the most general version, such terms are added to each function $g_m(t)$, leading to multivariate random effects.

In practice, a simpler model restricted to random intercept terms may suffice.[12] Such a restricted model would assume heterogeneity on the proportional part but not on the time-varying part of the hazard functions. In this case, the linear predictor becomes

$$\eta_{jkt} = \alpha_{0j} + \theta_{0k} + u_{jk} + \sum_{m=1}^{M} (\alpha_{mj} + \theta_{mk}) g_m(t),$$

where u_{jk} is a random intercept term with $E(u_{jk})$ = 0 to model between-trial heterogeneity. Different assumptions are possible to model heterogeneity, the simplest being independent normally distributed random effects with (same) variance $Var(u_{jk}) = \sigma^2$. The Supplementary Material (found at https://doi.org/10.1016/j.jval.2020.03.010) provides R and SAS code to fit such random-effects models.

Figure 1. Kaplan-Meier curves for the example data.



More general random-effects models and the relationship between different model parameterizations have been discussed (for NMA models without time-varying terms) in the aforementioned literature.[15,21,27] These results could be extended to our case, although this is beyond the scope of this work.

One should note that maximum likelihood estimation can be problematic in situations such as this, in which the number of parameters increases with the number of studies. Recent work shows that satisfactory results are achieved when using penalized quasi-likelihood/pseudo-likelihood methods.[27,28]

In what follows, we will concentrate on fixed-effect NMA models because our focus is on modeling the underlying time-varying hazard ratios.

Model Fitting, Model Building, and Model Selection

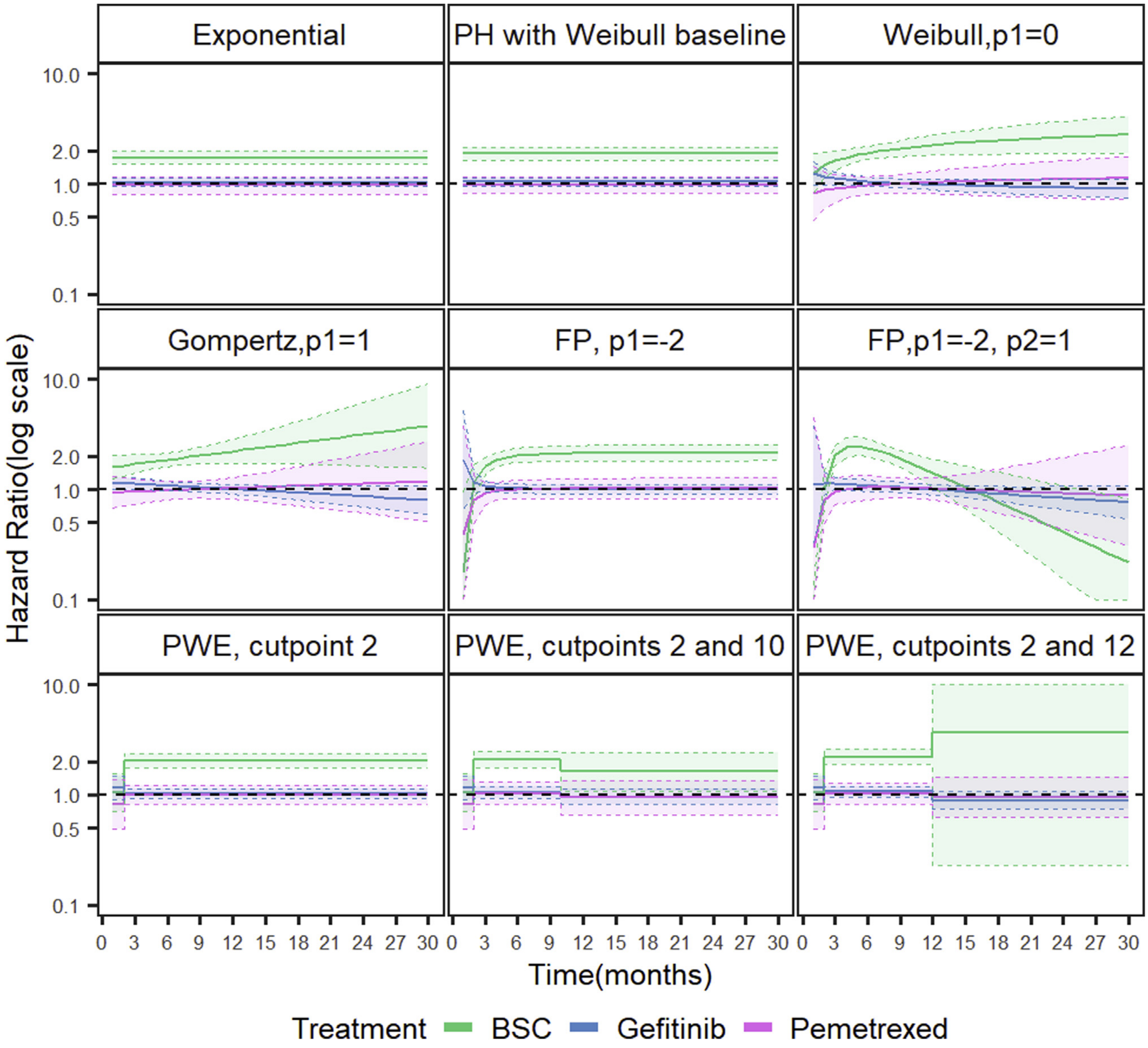
The models presented earlier can be fitted in a Bayesian or a frequentist framework. In the former case, the model needs be completed with suitable priors. For the frequentist analysis, standard programs such as glm() in R or PROC GLIMMIX in SAS may be used.

If flat priors were used, the fixed-effect NMA model would lead to estimates in close agreement between the Bayesian and the frequentist fits, and also the (posterior mean) deviance information criterion (DIC) from the Bayesian fit should be close to the (frequentist) Akaike information criterion (AIC).

Bayesian analysis using the random-effects model would require careful specification of the prior for between-study heterogeneity. Suitable priors for the random effects in (network) meta-analysis have been discussed elsewhere.[29-35]

Time-varying hazard ratio models are much more flexible than traditional TTE NMAs based on (single) hazard ratio estimates. Such flexibility comes at some risk of overfitting and requires careful model building and model selection strategies (see also chapter 10 of Dias et al[35]). Model selection should

Figure 2. Hazard ratios (other treatments vs docetaxel) over time (frequentist fixed-effect models).



consider model fit but also other aspects. We suggest analysts assess their models by inspecting model fit statistics (such as AIC or DIC), correlations between the multidimensional treatment effect estimates, visual inspection of model predictions against observed data, and the clinical plausibility of model extrapolations. Because the GLM approach allows the exploration of a large number of models very quickly, we propose a 2-step strategy for model building. In the first step, a large set of candidate models are fitted in a fixed-effect frequentist framework. This step serves to identify the "best" model(s), which should perform well on all criteria listed earlier. For the second step, the best structure is put into the Bayesian framework, in which both fixed- and random-effect(s) versions can be explored (the Bayesian implementation follows closely the example in Jansen[12]).

In cases in which overfitting is a concern, simplified models may be useful where the higher-order time-varying treatment effects θ_{mk} are set to zero, whereas the corresponding study effects α_{mj} are kept in the model. This allows flexibility for the baseline model but reduces the risk of overfitting by putting stronger restrictions on the treatment effect (including the special case of PH as discussed previously).

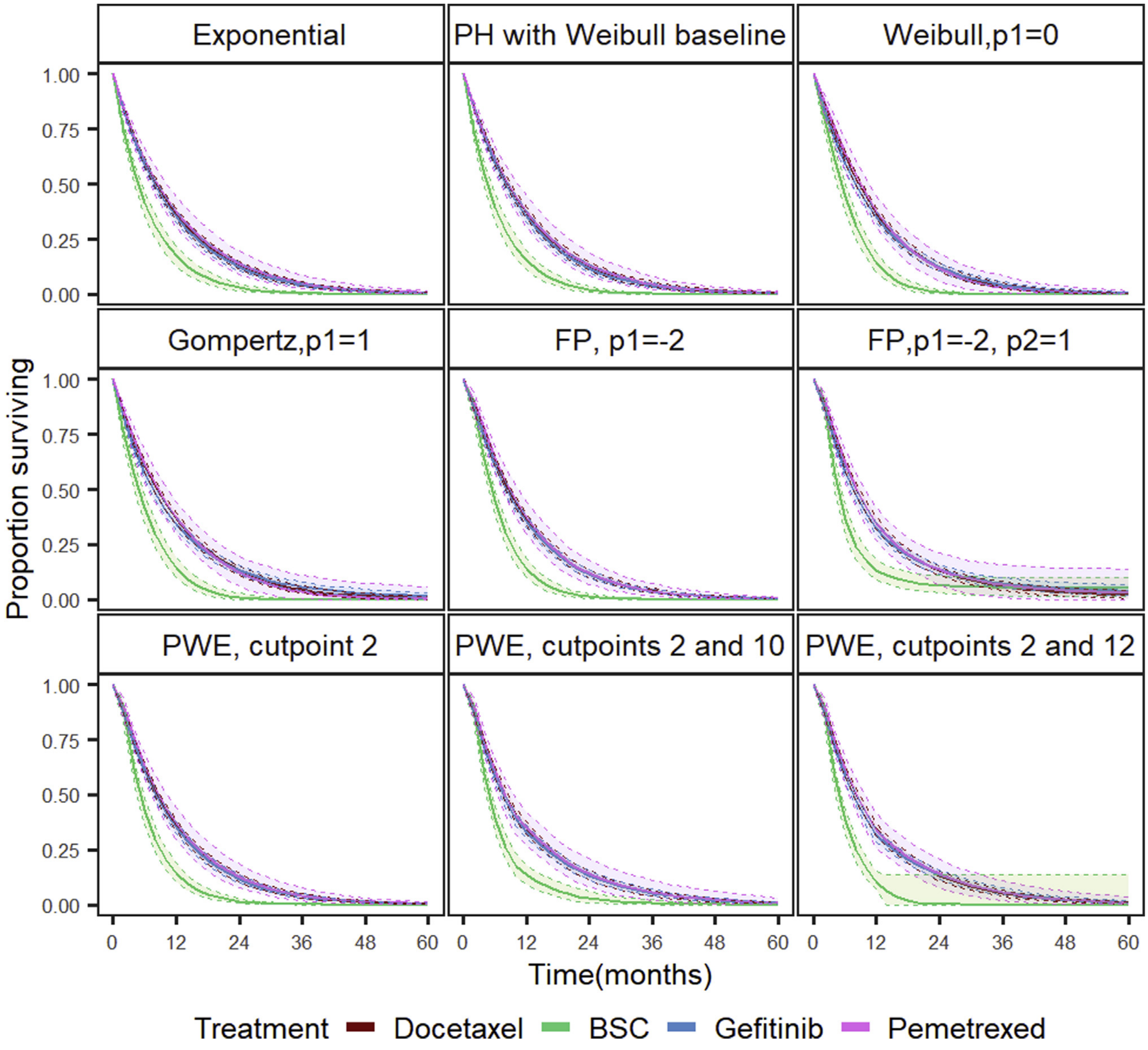
Consistency Assessment

The ANOVA-like parameterization allows for consistency assessment by the introduction of trial type and trial type by treatment interaction terms, where trial type refers to the set of treatments being compared in a given study.[22] In principle, our model could be extended by adding such interactions on all relevant terms (eg, g_0, g_1, g_2).

Illustrative Example

In the next section, we will apply the presented models to the example data from Jansen.[12] The example consists of survival data

Figure 3. Predicted survival, Kim (2008)[36] as baseline survival (frequentist fixed-effect models).



from 7 studies and 4 treatments (docetaxel, pemetrexed, gefitinib, and best supportive care [BSC]; Figure 1).

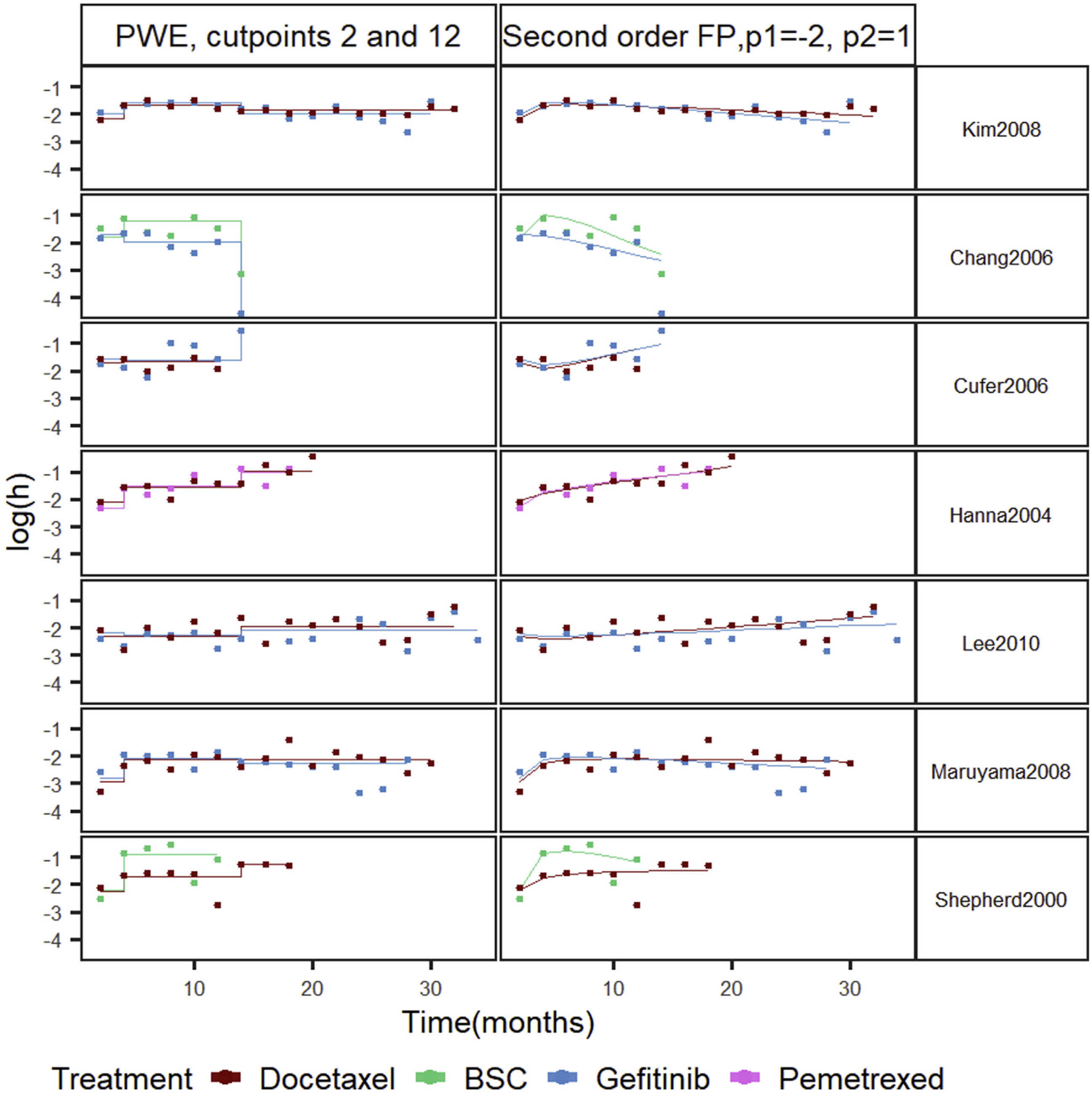
Results

We fit the first- and second-order fractional polynomial and PWE fixed-effect models using ANOVA parametrization in a frequentist framework. For fractional polynomials, we included all models presented in the original publication by Jansen.[12] For PWE models, we explored models with 1 or 2 cut points in the observed time range. For all fractional polynomial models, the AIC were similar to the DIC reported in Jansen[12] and also the parameter estimates matched. Results for all models can be produced with the R-code given in the Supplementary Material on the journal webpage.

Table 1 presents the AIC from a set of selected fractional polynomial and PWE models. The fractional polynomial model with the smallest AIC is the model with $p_1 = -2$ and $p_2 = 1$ for t^{p_m}. From the PWE models, the smallest AIC has the model with two cut points, one at 2 months and one at 12 months. The AIC for the best PWE model is 853.4 and for the best fractional polynomial model 866.5. If the model selection would be made purely based on AIC (or DIC in Bayesian framework), the PWE model with cut points at 2 and 12 months would be selected. Figure 2 shows the hazard ratios over time for the 9 models presented in Table 1. For all models, the comparison of pemetrexed versus docetaxel and gefitinib versus docetaxel produce similar shapes for hazard ratio over time. For the comparison of BSC versus docetaxel, the shape of the hazard ratio over time varies between different models. For the best model based on AIC, the PWE with two cut points, the hazard ratio versus docetaxel increased over time, whereas for the best fractional polynomial model, the hazard ratio initially increased but then decreased over time.

Figure 3 presents the predicted survival curves for the selected models using Kim (2008)[36] as the baseline study. Although the

Figure 4. Predicted versus observed data on log-hazard scale for the piecewise exponential and fractional polynomial models with lowest Akaike information criterion (frequentist fixed-effect models).



hazard ratios have very different shapes for the best fractional polynomial and PWE model, the predicted survival from both models are similar. It should be noted that the choice of baseline study can markedly affect the absolute survival estimates (whichever approach is used, the glm framework proposed here or the original method by Jansen). Supplementary Figure 1 shows the predicted survivor functions from the second-order fractional polynomial model obtained for each study in the data set selected as baseline. In practice, the choice of baseline from which to derive absolute outcomes requires special care.[35] In some cases, one study in the network may be a natural choice (as, for example, a pivotal trial for a new compound of interest). In other cases, an average baseline estimate from the studies in the network may be more suitable. Finally, a baseline estimate based on external data may be most appropriate in special circumstances too.

Figure 4 shows the observed data versus the predicted for the models with lowest AICs. The data are presented on the linear predictor scale. For Chang (2006)[37] and Cufer (2006),[38] there are only single data points in the last segment of the PWE model, and therefore the predicted value will match exactly the observed.

For comparison, we also fitted the second-order fractional polynomial model and one PWE model in a Bayesian framework using the baseline-contrast parameterization. Results were very similar between the frequentist and Bayesian fits (Table 2).

Table 2. Parameter estimates for the selected (fixed-effect) models.

		Second-order fractional polynomial				
		Frequentist glm fit			Bayesian baseline contrast fit	
		Estimate	95% confidence interval		Posterior median	95% credible interval
BSC vs docetaxel	θ_{0B}	1.668	(1.127-2.21)	d_{0BA}	1.672	(1.133-2.212)
Gefitinib vs docetaxel	θ_{0C}	0.172	(−0.061 to 0.405)	d_{0CA}	0.172	(−0.063 to 0.403)
Pemetrexed vs docetaxel	θ_{0D}	0.127	(−0.454 to 0.707)	d_{0DA}	0.126	(−0.455 to 0.71)
BSC vs docetaxel	θ_{1B}	−5.836	(−8.479 to −3.193)	d_{1BA}	−5.866	(−8.536 to −3.297)
Gefitinib vs docetaxel	θ_{1C}	−0.055	(−1.445 to 1.335)	d_{1CA}	−0.056	(−1.44 to 1.337)
Pemetrexed vs docetaxel	θ_{1D}	−1.313	(−4.404 to 1.779)	d_{1DA}	−1.312	(−4.445 to 1.735)
BSC vs docetaxel	θ_{2B}	−0.106	(−0.167 to −0.045)	d_{2BA}	−0.107	(−0.168 to −0.047)
Gefitinib vs docetaxel	θ_{2C}	−0.015	(−0.032 to 0.003)	d_{2CA}	−0.015	(−0.032 to 0.003)
Pemetrexed vs docetaxel	θ_{2D}	−0.008	(−0.06 to 0.044)	d_{2DA}	−0.008	(−0.061 to 0.043)

		Piecewise exponential with 2 cut points, 2 and 12 months				
		Frequentist glm fit			Bayesian baseline contrast fit	
		Estimate	95% confidence interval		Posterior median	95% credible interval
BSC vs docetaxel	θ_{0B}	0.045	(−0.35 to 0.44)	d_{0BA}	0.042	(−0.374 to 0.436)
Gefitinib vs docetaxel	θ_{0C}	0.144	(−0.098 to 0.385)	d_{0CA}	0.147	(−0.092 to 0.39)
Pemetrexed vs docetaxel	θ_{0D}	−0.194	(−0.703 to 0.315)	d_{0DA}	−0.195	(−0.723 to 0.317)
BSC vs docetaxel	θ_{1B}	0.749	(0.323 to 1.175)	d_{1BA}	0.752	(0.326-1.197)
Gefitinib vs docetaxel	θ_{1C}	−0.082	(−0.348 to 0.185)	d_{1CA}	−0.085	(−0.351 to 0.178)
Pemetrexed vs docetaxel	θ_{1D}	0.222	(−0.331 to 0.775)	d_{1DA}	0.222	(−0.332 to 0.797)
BSC vs docetaxel	θ_{2B}	1.254	(−1.552 to 4.061)	d_{2BA}	1.272	(−2.421 to 4.918)
Gefitinib vs docetaxel	θ_{2C}	−0.257	(−0.567 to 0.053)	d_{2CA}	−0.262	(−0.572 to 0.047)
Pemetrexed vs docetaxel	θ_{2D}	0.138	(−0.52 to 0.797)	d_{2DA}	0.139	(−0.53 to 0.809)

BSC indicates best supportive care.

Differences between the 2 frameworks would be mainly expected in the case of random-effects models with few degrees of freedom to estimate heterogeneity.

Discussion

TTE NMAs with time-varying hazard ratios can be fitted efficiently if the underlying log-hazard curves decompose as linear models. We have shown that both the now-prominent fractional polynomial models and the extremely flexible PWE models fall within this class. Via ANOVA parameterization of NMA, parameter estimation for these models is achieved via standard GLM routines. The frequentist fixed-effect version of these NMA models can serve to explore a large set of candidate structures rapidly. Analysts can therefore try out many hazard ratio shapes, compare model fit, and explore model properties and derived parameters. The best model from this step can then be analyzed more thoroughly.

Our fitting and model-building scheme works with grouped survival data, similar to the work in which fractional polynomial NMAs were introduced.[12] This is a limitation because in principle, the underlying continuous TTE data should be more informative than a discretized version. Also, all assumptions needed for NMA apply. Another limitation relates to the flexibility to use many different models itself. Although complex models with time-varying hazard ratios are appealing conceptually, they may be difficult to interpret and communicate. Fractional polynomials need upfront selection of order and exponents, and PWE models require selection of the number and location of cut points. In rare cases with a particularly large evidence base, structural unknowns such as the cut points might be estimated from the data. More often, clinical expertise may guide such choices. A structured framework for selecting these model "constants" is currently lacking and remains an area for future research. Our approach should not be taken as an invitation to fit blindly a large set of models without conceptual underpinning ("more" is not necessarily "better").

The ability to fit complex TTE NMA models with standard GLM routines is certainly a strength of our approach. Such routines exist in all major statistical packages, are well developed and robust, and fitting is usually extremely fast. Practical experience shows that, in particular, the fractional polynomial models can be time-consuming to fit in a Bayesian framework.[39] Even for the simple example used here, some models take several minutes to converge on a standard laptop. Although the frequentist result is certainly less informative than a full posterior, fitting a large set of candidate models in less than a second provides a very useful tool to the practitioner.

Our approach breaks down complexity of TTE NMAs with time-varying hazard ratios by first focusing on model structure, for which a frequentist fixed-effect framework is appropriate. More detailed analysis including modelling of heterogeneity (via random effects) and capturing and propagation of uncertainty (via Bayesian methods) can follow in a second step. Similar guidance has been given by Gelman and Hill[40] for multilevel hierarchical model building. This acknowledges that structural model selection goes beyond simple criteria of fit, such as AIC or DIC. Additional factors such as biological knowledge and plausibility of extrapolations matter greatly as well.[35] NMA results often feed into economic evaluations, where Bayesian methods have proven particularly useful to support decision-making.[41-43]

We did not explore the full space of TTE models for which our approach applies. The ANOVA-based parameterization works, in principle, whenever the log-hazard is a linear function of the model parameters. For example, spline-based log-hazards could also be used. Such models would closely resemble Royston-Parmar models, in which the log-cumulative hazard is expressed as a spline function.[44] It remains an open research question how far beyond fractional polynomials and PWE models the approach presented here can reach. We also focused on the fixed-effect NMA case, although random-effects NMAs can be expressed via the ANOVA parameterization and generalized linear mixed model (GLMM) fitting routines are well established. However, the baseline-contrast parameterization and the arm-level ANOVA parameterization can lead to different interpretations of the random effects.[21] Appropriate restrictions on the (in this case multivariate) random effects would be required. This can be handled using residual pseudo-likelihood as shown in Piepho et al.[27] Finally, we have only briefly touched upon the (important) topic of inconsistency assessment. The model could be extended to include type × treatment interactions to assess inconsistency. However, such an assessment at the level of each time-varying component might be difficult to interpret in practice. Optimal strategies for testing inconsistency with complex models represent an area for future research.

Our work extends the fundamental paper by Jansen[12] in which fractional polynomial NMAs were introduced. Although conceptually appealing, these models proved difficult to use in practice. Our approach makes a broad set of complex TTE models, including fractional polynomials, readily available for practitioners. Also, our approach fosters a holistic view on model building as typically needed in medical decision making. NMA often provides comparative effectiveness estimates as a basis to subsequent health economic evaluation. This requires thorough evaluation of model properties beyond statistical fit. Our approach allows exploring meaningful candidate model structures more efficiently. This will help analysts to better investigate, understand, and communicate the properties of time-varying hazard ratio models. Hopefully, this will ultimately result in better decision making when nonproportional hazards TTE data are part of the evidence base.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2020.03.010.

Article and Author Information

Accepted for Publication: March 22, 2020

Published Online: July 19, 2020

doi: https://doi.org/10.1016/j.jval.2020.03.010

Author Affiliations: Novartis Pharma AG, Basel, Switzerland (Wiksten); Institute of Health and Wellbeing, University of Glasgow, Glasgow, Scotland (Hawkins); Biostatistics Unit, University of Hohenheim, Stuttgart, Germany (Piepho); F. Hoffmann-La Roche AG, Basel, Switzerland (Gsteiger).

Author Contributions: *Concept and design:* Wiksten, Gsteiger.
Acquisition of data: Wiksten, Gsteiger.
Analysis and interpretations of data: Wiksten, Hawkins, Gsteiger.
Drafting of the manuscript: Wiksten, Hawkins, Piepho, Gsteiger.
Critical revision of the paper for important intellectual content: Hawkins, Piepho, Gsteiger.
Statistical analysis: Wiksten, Piepho, Gsteiger.
Supervision: Piepho, Gsteiger.

Conflict of Interest Disclosures: Dr Wiksten reported holding shares in Roche Ltd and NNovartis Pharma AG outside the submitted work. Dr Hawkins reported receiving personal fees from Roche Ltd outside the submitted work. No other disclosures were reported.

REFERENCES

1. Royston P, Parmar MKB. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Med Res Methodol*. 2016;16(16):1–13.
2. Freeman SC, Carpenter JR. Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models. *Res Synth Methods*. 2017;8(4):451–464.
3. Chen T-T. Statistical issues and challenges in immuno-oncology. *J Immunother Cancer*. 2013;1(18):1–9.
4. Cortés-Martínez J, Gómez-Mateu M, Kim K, Gómez-Melis G. Non-constant hazard ratios in randomized controlled trials with composite endpoints. *ArXiv190710976 Stat*. 2019.
5. Huang B. Some statistical considerations in the clinical development of cancer immunotherapies. *Pharm Stat*. 2018;17(1):49–60.
6. Rittmeyer A, Barlesi F, Waterkamp D, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet*. 2017; 389(10066):255–265.
7. Bellmunt J, de Wit R, Vaughn DJ, et al. Pembrolizumab as second-line therapy for advanced urothelial carcinoma. *N Engl J Med*. 2017;376(11):1015–1026.
8. Motzer RJ, Tannir NM, McDermott DF, et al. Nivolumab plus ipilimumab versus sunitinib in advanced renal-cell carcinoma. *N Engl J Med*. 2018;378(14):1277–1290.
9. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2 a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making*. 2013;33:607–617.
10. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol*. 2013;13(152):1–15.
11. Wei LJ. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Stat Med*. 1992;11(14):1871–1879.
12. Jansen JP. Network meta-analysis of survival data with fractional polynomials. *BMC Med Res Methodol*. 2011;11(61):1–14.
13. Jackson C, Stevens J, Ren S, et al. Extrapolating survival from randomized trials using external data: a review of methods. *Med Decis Making*. 2017;37(4):377–390.
14. Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Stat Med*. 2007;26(20):3681–3699.
15. Piepho HP, Williams ER, Madden LV. The use of two-way linear mixed models in multitreatment meta-analysis. *Biometrics*. 2012;68(4):1269–1277.
16. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2017.
17. StataCorp. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC; 2017.
18. Vickers AD, Winfree KB, Cuyun Carter G, et al. Relative efficacy of interventions in the treatment of second-line non-small cell lung cancer: a systematic review and network meta-analysis. *BMC Cancer*. 2019;19(353):1–16.
19. Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12(9):1–13.
20. Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*. 1978;34(1):57–67.
21. Hawkins N, Scott DA, Woods B. 'Arm-based' parameterization for network meta-analysis. *Res Synth Methods*. 2016;7(3):306–313.
22. Piepho H-P. Network-meta analysis made easy: detection of inconsistency using factorial analysis-of-variance models. *BMC Med Res Methodol*. 2014;14(61):1–9.
23. Ouwens MJNM, Philips Z, Jansen JP. Network meta-analysis of parametric survival curves. *Res Synth Methods*. 2010;1(3-4):258–271.
24. Hong H, Fu H, Price KL, Carlin BP. Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment. *Stat Med*. 2015;34(20):2794–2819.

25. Dias S, Ades AE. Absolute or relative effects? Arm-based synthesis of trial data. *Res Synth Methods*. 2016;7(1):23–28.
26. Lin L, Zhang J, Hodges JS, Chu H. Performing arm-based network meta-analysis in R with the pcnetmeta package. *J Stat Softw*. 2017;80(5):1–25.
27. Piepho H-P, Madden LV, Roger J, Payne R, Williams ER. Estimating the variance for heterogeneity in arm-based network meta-analysis. *Pharm Stat*. 2018;17(3):264–277.
28. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat Med*. 2018;37(7):1059–1085.
29. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal*. 2006;1(3):515–534.
30. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med*. 2005;24(15):2401–2428.
31. Pullenayegum EM. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Stat Med*. 2011;30(26):3082–3094.
32. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol*. 2012;41(3):818–827.
33. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med*. 2015;34(6):984–998.
34. Turner RM, Domínguez-Islas CP, Jackson D, Rhodes KM, White IR. Incorporating external evidence on between-trial heterogeneity in network meta-analysis. *Stat Med*. 2019;38(8):1321–1335.
35. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. *Network Meta-Analysis for Decision Making*. Hoboken, NJ: Wiley; 2018.
36. Kim ES, Hirsh V, Mok T, et al. Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial. *Lancet*. 2008;372(9652):1809–1818.
37. Chang A, Parikh P, Thongprasert S, et al. Gefitinib (IRESSA) in patients of Asian origin with refractory advanced non-small cell lung cancer: subset analysis from the ISEL study. *J Thorac Oncol*. 2006;1(8):847–855.
38. Cufer T, Vrdoljak E, Gaafar R, Erensoy I, Pemberton K, SIGN Study Group. Phase II, open-label, randomized study (SIGN) of single-agent gefitinib (IRESSA) or docetaxel as second-line therapy in patients with advanced (stage IIIb or IV) non-small-cell lung cancer. *Anticancer Drugs*. 2006;17(4):401–409.
39. Laliman VA, Pacou M, Gauthier A. Fractional polynomial NMA models for survival analyses: results from a simulation study. *Value Health*. 2017;20:A770–A771.
40. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press; 2006.
41. Baio G. *Bayesian Methods in Health Economics*. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2013.
42. Briggs A, Claxton K, Sculpher M. *Decision Modelling for Health Economic Evaluation*. Oxford, UK: Oxford University Press; 2006.
43. Welton NJ, Sutton AJ, Cooper NJ, Abrams KR, Ades AE. *Evidence Synthesis for Decision Making in Healthcare*. Chichester, UK: John Wiley & Sons; 2012.
44. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med*. 2002;21(15):2175–2197.

5 Hartung–Knapp method is not always conservative compared with fixed-effect meta-analysis

Research Article

Statistics in Medicine

Received 9 April 2015, Accepted 3 January 2016 Published online 4 February 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6879

Hartung–Knapp method is not always conservative compared with fixed-effect meta-analysis

Anna Wiksten,[a*†] Gerta Rücker[b] and Guido Schwarzer[b]

A widely used method in classic random-effects meta-analysis is the DerSimonian–Laird method. An alternative meta-analytical approach is the Hartung–Knapp method. This article reports results of an empirical comparison and a simulation study of these two methods and presents corresponding analytical results. For the empirical evaluation, we took 157 meta-analyses with binary outcomes, analysed each one using both methods and performed a comparison of the results based on treatment estimates, standard errors and associated *P*-values. In several simulation scenarios, we systematically evaluated coverage probabilities and confidence interval lengths. Generally, results are more conservative with the Hartung–Knapp method, giving wider confidence intervals and larger *P*-values for the overall treatment effect. However, in some meta-analyses with very homogeneous individual treatment results, the Hartung–Knapp method yields narrower confidence intervals and smaller *P*-values than the classic random-effects method, which in this situation, actually reduces to a fixed-effect meta-analysis. Therefore, it is recommended to conduct a sensitivity analysis based on the fixed-effect model instead of solely relying on the result of the Hartung–Knapp random-effects meta-analysis. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: meta-analysis; Hartung–Knapp method; DerSimonian-Laird method; empirical evaluation

1. Introduction

A common problem in meta-analysis is the decision between fixed-effect and random-effects model. In a fixed-effect model, the true treatment effect is assumed to be the same in all studies, and the observed treatment effects vary only because of random errors inherent in each study. This assumption is often not reasonable because studies vary by several items, for example, inclusion criteria, geographical location or the implementation of interventions, which typically leads to greater heterogeneity between-study specific treatment estimates. Treatment effects are called heterogeneous if the observed treatment effects vary more than we would expect by chance. In this case, the use of a random-effects model is commonly recommended.

The DerSimonian–Laird method [1] is the most widely used method to estimate the between-study variance in systematic reviews and meta-analyses. For example, this is the only method available in Review Manager 5 [2], the software programme of the Cochrane Collaboration to prepare and maintain Cochrane reviews. Several other methods to estimate the between-study variance have been proposed [3, 4].

Hartung and Knapp [5–7] introduced an alternative meta-analytical approach based on a different variance estimator in the random-effects model. Sidik and Jonkman [8] independently proposed the same method a couple of years after the initial publication by Hartung [5]. Accordingly, this approach is sometimes called the Hartung–Knapp–Sidik–Jonkman method [9]. It has been shown in simulations [7, 9, 10] that a test based on the Hartung–Knapp modification holds the prespecified significance level much better than tests based on the classic fixed-effect and random-effects model.

[a] *Statistical Methodology, Development, Novartis Pharma AG, Basel, Switzerland*
[b] *Institute for Medical Biometry and Statistics, Medical Center - University of Freiburg, Freiburg, Germany*
**Correspondence to: Anna Wiksten, Statistical Methodology, Development, Novartis Pharma AG, CH-4002 Basel,Switzerland.*
[†] *E-mail: a.s.wiksten@gmail.com*

In recent publications, the use of the classic DerSimonian–Laird method has been condemned, and the Hartung–Knapp method was suggested as possible alternative both in simulations [9] and in empirical evaluations either based on 689 meta-analyses from Cochrane reviews [9] or a single classic example [11].

Note, both classic random-effects method and Hartung–Knapp modification use the same pooled treatment-effect estimate; however, formulae for standard errors are different, and quantiles of the standard normal distribution (classic method) and *t*-distribution (Hartung–Knapp approach) are used in order to construct confidence intervals and calculate *P*-values. These differences affect the length of confidence intervals and *P*-values.

The aim of this article is to compare the Hartung–Knapp method and standard DerSimonian–Laird method empirically in a set of 157 meta-analyses and to support the findings with analytical and simulation results. We use the set of 157 meta-analyses with binary outcomes provided by Jüni [12] and described by us in [13]. The number of studies in each meta-analysis ranges from 4 to 66 (median 8), and the number of patients in the component studies ranges from 2 to 15280 (median 103). For each study within each meta-analysis, data are available on the number of patients and number of events in each treatment group.

The plan for the rest of the article is as follows. Section 2 reviews the compared methodologies. In Section 3, we briefly describe our empirical evaluation and report the results. In Section 4, we present results from a simulation study comparing the two methods, and in Section 5, we discuss our findings.

2. Review of the compared methods

In this section, we briefly review the fixed-effect model and two random-effects approaches, namely the classic DerSimonian–Laird and Hartung–Knapp methods. All methods require from each included study an estimated treatment effect and its standard error as input. In meta-analyses with binary outcomes, the log odds ratio or the log risk ratio is typically used as measure of treatment effect.

2.1. *Fixed-effect meta-analysis*

The general fixed-effect model is

$$\hat{\theta}_k = \theta + \sigma_k \epsilon_k, \quad \epsilon_k \sim N(0,1) \tag{1}$$

where $\hat{\theta}_k$ is the observed treatment effect in study k, $k = 1, \ldots, K$, σ_k^2 is the study specific variance and θ is the unknown true treatment effect, which is common for all studies. The fixed-effect estimate of θ is denoted by $\hat{\theta}_F$. Given $\left(\hat{\theta}_k, \hat{\sigma}_k\right)$, $k = 1, \ldots, K$, the maximum likelihood estimate under model (1) is

$$\hat{\theta}_F = \frac{\sum_{k=1}^{K} \hat{\theta}_k / \hat{\sigma}_k^2}{\sum_{k=1}^{K} 1/\hat{\sigma}_k^2} = \frac{\sum_{k=1}^{K} w_k \cdot \hat{\theta}_k}{\sum_{k=1}^{K} w_k} \tag{2}$$

with weights $w_k = 1/\hat{\sigma}_k^2$.

The variance of $\hat{\theta}_F$ is estimated by

$$\widehat{\mathrm{Var}}\left(\hat{\theta}_F\right) = \frac{1}{\sum_{k=1}^{K} w_k}. \tag{3}$$

A (1-α) confidence interval for $\hat{\theta}_F$ can be calculated by

$$\hat{\theta}_F \pm z_{1-\frac{\alpha}{2}} \times \text{S.E.}\left(\hat{\theta}_F\right) \tag{4}$$

with standard error $\text{S.E.}\left(\hat{\theta}_F\right) = \sqrt{\widehat{\mathrm{Var}}\left(\hat{\theta}_F\right)}$ and $z_{1-\frac{\alpha}{2}}$ denoting the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution. A corresponding test for an overall treatment effect can be constructed using $\hat{\theta}_F \big/ \text{S.E.}\left(\hat{\theta}_F\right)$ as test statistic.

2.2. Classic random-effects meta-analysis using DerSimonian–Laird method

In contrast to a fixed-effect model, a random-effects model allows that the underlying true treatment effects in individual studies vary, typically according to a normal distribution

$$\hat{\theta}_k = \theta_k + \sigma_k \epsilon_k, \quad \epsilon_k \sim N(0,1);\ \theta_k \sim N\left(\theta, \tau^2\right). \tag{5}$$

The fixed-effect model is a special case of the random effects model when the between-study variance $\tau^2 = 0$. The between-study variance τ^2 is an additional parameter that has to be estimated in the random-effects model.

Several methods have been proposed for the estimation of τ^2 [3,4], of which so far the most popular is the DerSimonian–Laird method [1]. It is a non-iterative, moment-based estimator for the between-study variance τ^2

$$\hat{\tau}^2 = \frac{Q - (K-1)}{\sum_{k=1}^K w_k - \frac{\sum_{k=1}^K w_k^2}{\sum_{k=1}^K w_k}} \tag{6}$$

where Q is the heterogeneity statistic given by $Q = \sum_{k=1}^K w_k \left(\hat{\theta}_k - \hat{\theta}_F\right)^2$ and $w_k = \widehat{\text{Var}}\left(\hat{\theta}_k\right)^{-1}$. By definition, a variance cannot have negative values, and therefore, the estimate $\hat{\tau}^2$ is set to zero if $Q < K-1$, which corresponds to using a fixed-effect model.

The random effects estimate $\hat{\theta}_R$ and its variance can be calculated as

$$\hat{\theta}_R = \frac{\sum_{k=1}^K w_k^* \cdot \hat{\theta}_k}{\sum_{k=1}^K w_k^*} \qquad \widehat{\text{Var}}\left(\hat{\theta}_R\right) = \frac{1}{\sum_{k=1}^K w_k^*} \tag{7}$$

with weights $w_k^* = 1/\left(\hat{\sigma}_k^2 + \hat{\tau}^2\right)$. Note, compared with the fixed-effect model, calculating an overall effect estimate will pay greater attention to effect estimates from smaller studies. We will come back to this point in the discussion.

A (1-α) confidence interval for $\hat{\theta}_R$ can be calculated by

$$\hat{\theta}_R \pm z_{1-\frac{\alpha}{2}} \times \text{S.E.}\left(\hat{\theta}_R\right) \tag{8}$$

with standard error $\text{S.E.}\left(\hat{\theta}_R\right) = \sqrt{\widehat{\text{Var}}\left(\hat{\theta}_R\right)}$ and $z_{1-\frac{\alpha}{2}}$ denoting the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution. A corresponding test for an overall treatment effect can be constructed using $\hat{\theta}_R / \text{S.E.}\left(\hat{\theta}_R\right)$ as test statistic.

2.3. Hartung–Knapp method

Hartung and Knapp [5–7] introduced a new meta-analysis method based on a refined variance estimator in the random-effects model.

The overall treatment estimate in the Hartung–Knapp method is the same as in the classic random-effects model; however, instead of using the variance estimate given in (7), Hartung and Knapp propose to use the following variance estimator for $\hat{\theta}_R$:

$$\widehat{\text{Var}}_{HK}\left(\hat{\theta}_R\right) = \frac{1}{K-1} \sum_{k=1}^K \frac{w_k^*}{w^*} \left(\hat{\theta}_k - \hat{\theta}_R\right)^2 \tag{9}$$

with weights $w_k^* = 1/\left(\hat{\sigma}_k^2 + \hat{\tau}^2\right)$ and $w^* = \sum_{k=1}^K w_k^*$. Note, the Hartung–Knapp method is based on the weights w_k^* and the estimate $\hat{\theta}_R$ from the classic random-effects model. Accordingly, an estimate of the between-study variance $\hat{\tau}^2$ is also needed for the Hartung–Knapp method. In principle, any method to estimate τ^2 [4] can be used in the Hartung–Knapp method. Here, we will use the DerSimonian–Laird estimate of the between-study variance.

Hartung [5] showed that

$$\frac{\hat{\theta}_R - \theta}{\text{S.E.}_{HK}\left(\hat{\theta}_R\right)}$$

with standard error $\text{S.E.}_{HK}\left(\hat{\theta}_R\right) = \sqrt{\widehat{\text{Var}}_{HK}\left(\hat{\theta}_R\right)}$ follows a t-distribution with $K-1$ degrees of freedom.

Accordingly, a (1-α) confidence interval for $\hat{\theta}_R$ based on the Hartung–Knapp method can be calculated by

$$\hat{\theta}_R \pm t_{K-1;1-\frac{\alpha}{2}} \times \text{S.E.}_{HK}\left(\hat{\theta}_R\right) \tag{10}$$

with $t_{K-1;1-\frac{\alpha}{2}}$ denoting the $1-\frac{\alpha}{2}$ quantile of the t-distribution with $K-1$ degrees of freedom. A corresponding test for an overall treatment effect can be constructed using $\hat{\theta}_R / \text{S.E.}_{HK}\left(\hat{\theta}_R\right)$ as test statistic.

It has been shown in simulations [9, 14] that a test based on the Hartung–Knapp modification holds the prespecified significance level much better than tests based on $\text{S.E.}\left(\hat{\theta}_F\right)$ and $\text{S.E.}\left(\hat{\theta}_R\right)$, respectively.

3. Empirical evaluation

To compare the methods, we performed a random-effects meta-analysis based on the DerSimonian–Laird and Hartung–Knapp method in all 157 data sets. For each data set, we calculated the overall effect estimate with 95% confidence interval and corresponding P-value for the test of an overall treatment effect. We also calculated the P-value of the test for heterogeneity based on Cochran's Q for each data set. We used R package meta [15] to conduct meta-analyses using both DerSimonian–Laird and Hartung–Knapp methods. These methods are also available in R package metafor [16], which implements additional methods to estimate the between-study variance.

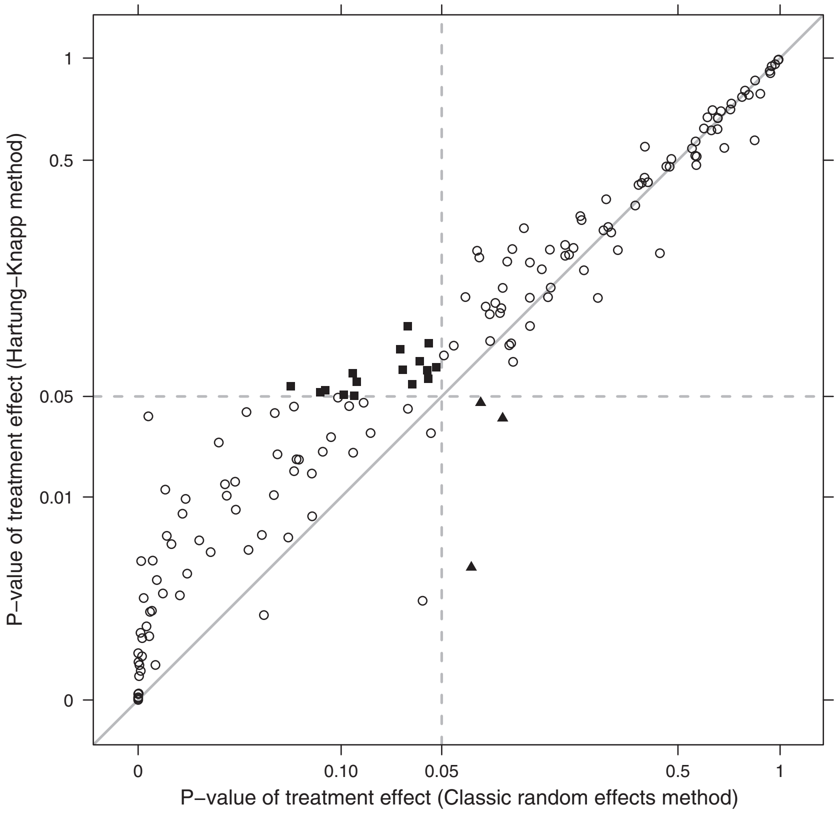


Figure 1. Comparison of treatment effect P-values between DerSimonian–Laird and Hartung–Knapp method.

Figure 1 shows the *P*-value for an overall treatment effect in the 157 meta-analyses, both for DerSimonian–Laird and Hartung–Knapp methods. Generally, we see a clear tendency – as expected from simulation studies [9, 14] – that *P*-values are greater with the Hartung–Knapp method. However, in several meta-analyses (33 out of 157), *P*-values are smaller with the Hartung–Knapp method than with the DerSimonian–Laird method. In Figure 1, we can see three meta-analyses (filled triangles) where the overall treatment effect is statistically significant with the Hartung–Knapp method, however, non-significant with the DerSimonian–Laird method.

We have a closer look at two extreme examples in order to investigate why the *P*-value of the Hartung–Knapp method is so markedly smaller. These two examples with five and seven studies have *P*-values for the DerSimonian–Laird method close to 0.05 and close to 0.001 for the Hartung–Knapp method.

3.0.1. Two illustrative examples. Figures 2 and 3 show the forest plots from the two most extreme meta-analyses. For the meta-analysis in Figure 2, the *P*-value for an overall treatment effect is 0.073 and 0.002 using the DerSimonian–Laird and Hartung–Knapp methods, respectively. In Figure 3, the respective *P*-values are 0.039 and 0.001. When we have a closer look at the forest plots, we see that in both meta-analyses, the estimated between-study variance $\hat{\tau}^2$ is zero, and the *P*-value for the heterogeneity test is very close to 1. In both examples, the heterogeneity statistic Q is far below its expected value. Accordingly, in these meta-analyses, the result for the DerSimonian–Laird method corresponds to the result from a fixed-effect meta-analysis.

3.0.2. Comparison of the methods by significance of heterogeneity test. Based on the findings in the two extreme meta-analyses, we wanted to further evaluate the difference in *P*-values for the test of an overall treatment effect based on the level of heterogeneity. Results from this evaluation are shown in Figure 4. Meta-analyses were divided into three groups according to the *P*-value of the heterogeneity test based on Cochran's Q.

Left and middle panels of Figure 4 clearly show that the *P*-value for an overall treatment effect is typically greater using the Hartung–Knapp method than using the DerSimonian–Laird method; most of the points (51 out 52 in the left panel, 66 out of 76 in the middle panel) lie above the identity line for *P*-values of the heterogeneity test below 0.7. This observation is in agreement with results from simulation studies [9, 14].

On the contrary, we see that the majority of points (22 out of 28) lie below the identity line in meta-analyses without indication of between-study heterogeneity (right panel in Figure 4, *P*-value of heterogeneity test above 0.7). In these 28 meta-analyses, the DerSimonian–Laird estimate of the between-study variance is always zero, which corresponds to using a fixed-effect meta-analysis.

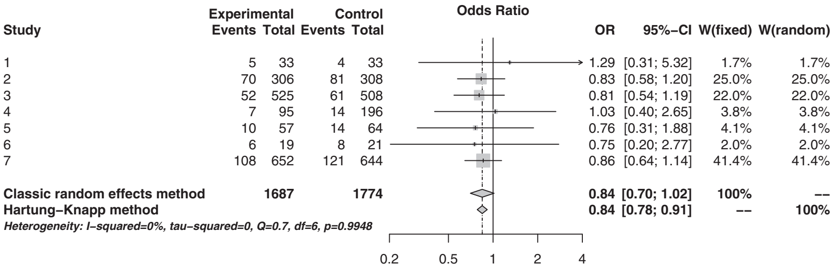


Figure 2. Forest plot for illustrative example 1.

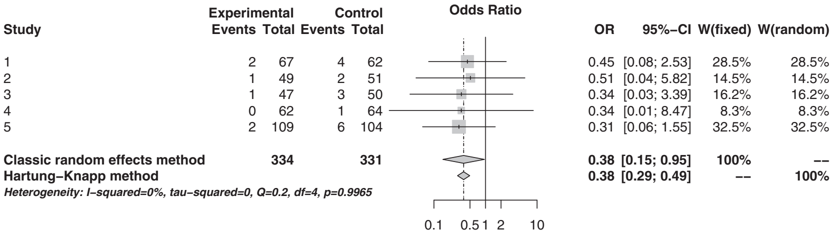


Figure 3. Forest plot for illustrative example 2.

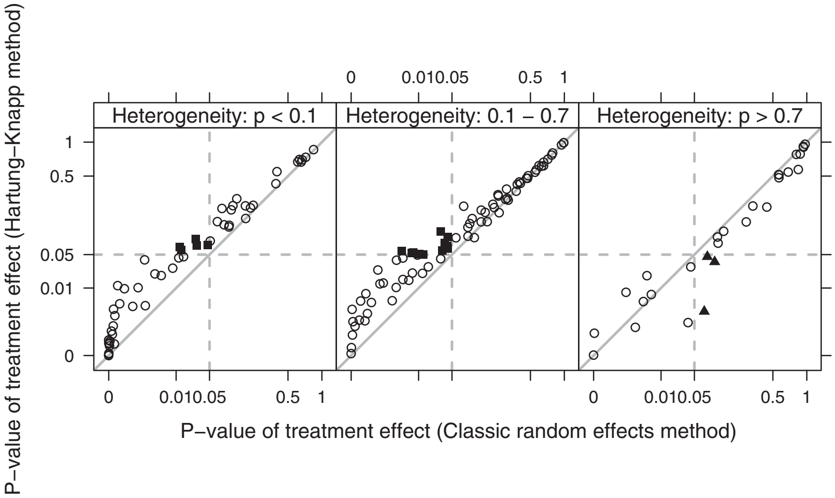


Figure 4. Comparison of treatment effect *P*-values between DerSimonian–Laird and Hartung–Knapp methods by the *P*-value of the heterogeneity test

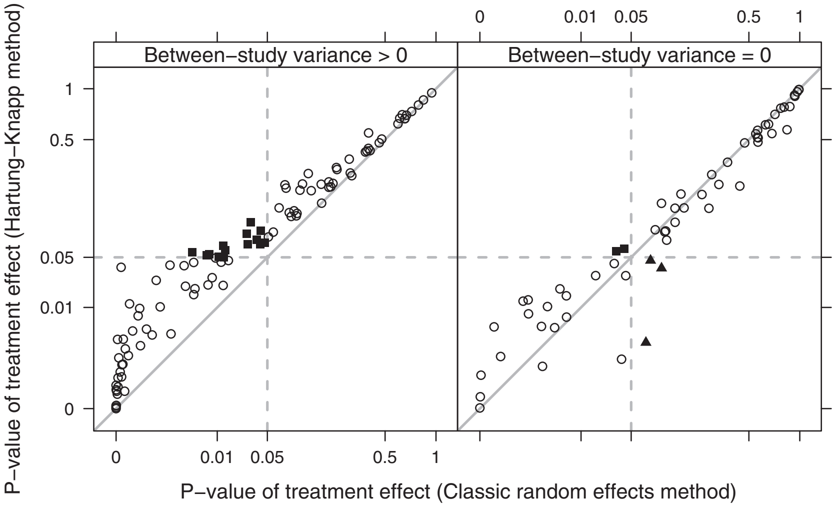


Figure 5. Comparison of treatment effect *P*-values between DerSimonian–Laird and Hartung–Knapp methods by $\hat{\tau}^2_{\mathrm{DSL}}$ values

3.0.3. Comparison of the methods by the between-study variance. The estimated between-study variance was zero in most of the cases (31 out of 33) if the DerSimonian–Laird method was conservative compared with the Hartung–Knapp method; actually, in the two meta-analyses with an estimated between-study variance greater than zero, the *P*-values for an overall treatment effect were very similar: $p = 0.1705$ versus $p = 0.1702$ and $p = 0.295$ versus $p = 0.281$, respectively. Figure 5 shows the comparison of the *P*-values in subgroups defined by the estimated between-study variance $\left(\hat{\tau}^2 > 0 \text{ vs } \hat{\tau}^2 = 0\right)$. In the left panel, we see that almost all points (97 out of 99) lie above the identity line if the estimated between-study variance is greater than zero, whereas in the right panel more points (31 out of 58) lie below the identity line meaning that the DerSimonian–Laird method is conservative compared with the Hartung–Knapp method.

3.0.4. Comparison of standard error of treatment effect by the between-study variance. In Figure 6, we compare the standard errors of the random effects treatment estimates in subgroups defined by the value of the between-study variance. We see that the standard error of the treatment effect is always smaller for the Hartung–Knapp method than for the DerSimonian–Laird method if the between-study variance is zero (right panel); as noted before, the *P*-value for the DerSimonian–Laird method is actually based on a fixed-effect model.

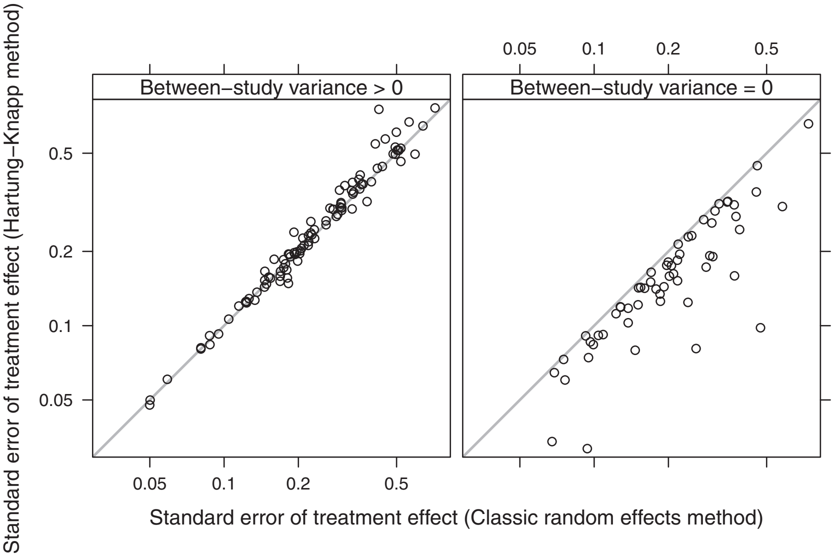


Figure 6. Comparison of standard errors of treatment effect between DerSimonian–Laird and Hartung–Knapp methods by $\hat{\tau}^2_{\text{DSL}}$ values

We provide a proof that the standard error of the treatment effect is always smaller or equal for the Hartung–Knapp method than for the DerSimonian–Laird method if the between-study variance estimate is zero. In this special case, we can rewrite Eq. 9 in the following way:

$$\begin{aligned}\widehat{\text{Var}}_{HK}(\hat{\theta}_R) &= \frac{1}{K-1}\sum_{k=1}^{K}\frac{w_k^*}{w^*}\left(\hat{\theta}_k-\hat{\theta}_R\right)^2\\ &= \frac{\sum_{k=1}^{K} w_k^*\left(\hat{\theta}_k-\hat{\theta}_R\right)^2}{K-1}\quad\frac{1}{\sum_{k=1}^{K} w_k^*}\\ &= \frac{\sum_{k=1}^{K} w_k\left(\hat{\theta}_k-\hat{\theta}_F\right)^2}{K-1}\quad\frac{1}{\sum_{k=1}^{K} w_k}\\ &= \frac{Q}{K-1}\quad\widehat{\text{Var}}\left(\hat{\theta}_F\right).\end{aligned} \tag{11}$$

In the third line, we replace w_k^* and $\hat{\theta}_R$ with w_k and $\hat{\theta}_F$, respectively, which is justified as $w_k^* = 1/\left(\hat{\sigma}_k^2+\hat{\tau}^2\right) = 1/\left(\hat{\sigma}_k^2+0\right) = 1/\hat{\sigma}_k^2 = w_k$; $\hat{\theta}_R = \hat{\theta}_F$ follows accordingly. In the fourth line, we notice that the nominator is equal to the heterogeneity statistic Q and that the estimated variance of the fixed-effect estimate is equal to the inverse of the sum of weights w_k.

For the DerSimonian–Laird method, the estimated between-study variance $\hat{\tau}^2$ is equal to zero if and only if $Q \leqslant (K-1)$ (Eq. 6). Therefore, the following relationship holds if the estimated between-study variance is zero as $Q/(K-1) \leqslant 1$:

$$\widehat{\text{Var}}_{HK}(\hat{\theta}_R) \leqslant \widehat{\text{Var}}\left(\hat{\theta}_F\right).$$

This relationship also holds for the standard error, that is, the square-root of the variance.

Although the standard error of the treatment effect is always smaller in this situation with the Hartung–Knapp method, the P-value is greater in roughly half of the cases (Figure 5, right panel). The explanation is that the Hartung–Knapp method uses a quantile of the t-distribution in the calculation of the P-value, whereas the DerSimonian–Laird method uses a quantile of the standard normal distribution, which is always smaller or equal than the quantile of a t-distribution.

When the estimated between-study variance is greater than zero, the estimated treatment-effect standard error by the Hartung–Knapp method is smaller than with the DerSimonian–Laird method in 31 of

99 cases. We will show, if the between-study variance τ^2 is known, that the probability of $\widehat{\mathrm{Var}}_{HK}\left(\hat{\theta}_R\right)$ being smaller than $\widehat{\mathrm{Var}}\left(\hat{\theta}_R\right)$ is always greater than 0.5. The formula for $\widehat{\mathrm{Var}}_{HK}\left(\hat{\theta}_R\right)$ is

$$\widehat{\mathrm{Var}}_{HK}\left(\hat{\theta}_R\right) = \frac{1}{K-1}\sum_{k=1}^{K}\frac{w_k^*}{w^*}\left(\hat{\theta}_k - \hat{\theta}_R\right)^2.$$

By multiplying both sides by w^* and $K-1$, we get

$$\begin{aligned} w^*(K-1)\,\widehat{\mathrm{Var}}_{HK}\left(\hat{\theta}_R\right) &= \sum_{k=1}^{K} w_k^*\left(\hat{\theta}_k - \hat{\theta}_R\right)^2 \\ &= \sum_{k=1}^{K}\frac{\left(\hat{\theta}_k - \hat{\theta}_R\right)^2}{\hat{\sigma}_k^2+\tau^2}. \end{aligned} \quad (12)$$

From Eq. 12, we can see that, given $\left(\hat{\theta}_k, \hat{\sigma}_k, \tau^2\right)$, $w^*(K-1)\,\widehat{\mathrm{Var}}_{HK}\left(\hat{\theta}_R\right)$ follows a χ^2-distribution with $K-1$ degrees of freedom. As the expected value of this distribution is $K-1$, the expected value of $\widehat{\mathrm{Var}}_{HK}\left(\hat{\theta}_R\right)$ is $1/w^* = \mathrm{Var}\left(\theta_R\right)$. Hence,

$$Pr\left(\widehat{\mathrm{Var}}_{HK}(\theta_R) < \widehat{\mathrm{Var}}\left(\theta_R\right)\right) = Pr\left(\chi^2_{K-1} < K-1\right).$$

Because the χ^2-distribution is skewed to the right, its expectation $(K-1)$ exceeds its median, and therefore,

$$Pr\left(\chi^2_{K-1} < K-1\right) > 0.5.$$

Therefore, the probability that $\widehat{\mathrm{Var}}_{HK}(\theta_R) < \widehat{\mathrm{Var}}\left(\theta_R\right)$ is always larger than 0.5, converging to 0.5 as the number of studies goes to ∞.

3.0.5. Confidence interval lengths. As shown in (11), the DerSimonian–Laird method collapses to the fixed-effect model, and the variance of the Hartung–Knapp method can be written as $\frac{Q}{K-1}\widehat{\mathrm{Var}}\left(\hat{\theta}_F\right)$, if $\hat{\tau}^2 = 0$. In this case, the length of the confidence interval for the DerSimonian–Laird method is

$$2 \times z_{1-\frac{\alpha}{2}} \times \mathrm{S.E.}\left(\hat{\theta}_F\right), \quad (13)$$

and for the Hartung–Knapp method,

$$2 \times t_{K-1;1-\frac{\alpha}{2}} \times \sqrt{\frac{Q}{K-1}}\mathrm{S.E.}\left(\hat{\theta}_F\right). \quad (14)$$

Following (13) and (14), the Hartung–Knapp method is anti-conservative if and only if

$$H^2 = \frac{Q}{K-1} < \left(\frac{z_{1-\frac{\alpha}{2}}}{t_{K-1;1-\frac{\alpha}{2}}}\right)^2 \quad (15)$$

where H^2 is a so-called scaling factor [14, 17].

4. Simulation study

We performed a simulation study to confirm the findings of the empirical evaluation and to further evaluate the properties of the two meta-analysis methods. We generated several simulation scenarios by varying the number of studies, K, the between-study variance, τ^2, and the average probability of an event in treatment and control group, p. Sample sizes for treatment groups were determined by randomly selecting studies from the set of 157 meta-analyses. We simulated the data from a binomial distribution and used log odds ratio and normal approximation for the analysis of individual studies in each meta-analysis.

We considered two simulation settings to evaluate different properties of the two approaches. In the first setting, we varied the number of studies in meta-analysis from $K = 2$ to 1000 in order to investigate small and large sample properties (with respect to the number of studies). In the second setting, we considered meta-analyses with $K = 5$, 10 and 50 studies in order to look in more detail on properties in typical meta-analyses ($K = 5$, 10). In both simulation settings, an average probability of $p = 0.1$ and 0.5 was considered. The study specific true treatment effect θ_k was simulated from a normal distribution with mean $\theta = 0$ (odds ratio of 1) and the following values for the between-study variance: $\tau^2 = 0$, 0.05, 0.1 and 0.2 (first setting) and $\tau^2 = 0$, 0.1 (second setting). For each combination of K, τ^2 and p, we conducted 10 000 (first setting)/100 000 (second setting) classic random effects and Hartung–Knapp meta-analyses using the DerSimonian–Laird method to estimate the between-study variance.

Results of the first simulation setting are summarised in Figure7 showing the relation between number of studies per meta-analysis (x-axis) and proportion of runs where the Hartung–Knapp confidence interval is shorter than the fixed-effect model confidence interval (y-axis). For $K = 2$, the confidence interval of the Hartung–Knapp method is shorter (anti-conservative) in approximately 10% of runs regardless of the underlying size of the between-study heterogeneity.

Under the fixed-effect model $\left(\tau^2 = 0\right)$, the proportion of anti-conservative Hartung–Knapp confidence intervals is (almost monotonically) increasing, which can be explained as follows: (i) an estimate $\hat{\tau}^2 = 0$ is rather likely under a fixed-effect model; (ii) if $\hat{\tau}^2 = 0$, standard error is always smaller for Hartung–Knapp approach than classic fixed-effect or random-effects meta-analysis (11); and (iii) influence of standard error on confidence-interval length is increasing with growing number of studies in meta-analysis as quantiles of t-distribution and z-distribution are getting more and more similar. For $p = 0.1$, the proportion of anti-conservative Hartung–Knapp confidence intervals is much higher than for $p = 0.5$. The probable reason for this phenomenon is that underdispersion of binomial responses is more likely in the case of rare events [18], and therefore, meta-analysis results are more homogeneous than expected under a fixed-effect model.

Under a random-effects model $\left(\tau^2 > 0\right)$, the proportion of anti-conservative Hartung–Knapp results increases for small number of studies (e.g. $K < 5$ for $\tau^2 = 0.2$ and $p = 0.1$); however, with increasing number of studies, the proportion decreases and approaches zero for very large numbers of studies; the decrease is faster for larger τ^2 and $p = 0.5$. This behaviour can be explained as follows: (i) an estimate $\hat{\tau}^2 = 0$ is more likely under a random-effects model for a small number of studies and small τ^2, and (ii) in general, the precision of $\hat{\tau}^2$ is very low for small number of studies resulting in larger variation in estimates and thus a larger proportion with $\hat{\tau}^2 = 0$. For very large numbers of studies, we extremely rarely observe runs with $\hat{\tau}^2 = 0$ under a random-effects model $\left(\tau^2 > 0\right)$.

In the second simulation setting, we calculated the observed H^2 for each simulation run and divided the 100 000 runs into eleven H^2 categories by using the following cutpoints: 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0. Note, if the scaling factor H^2 is smaller or equal to 1, the classic random-effects model collapses to the fixed-effect model as $\hat{\tau}^2 = 0$.

In Figure 8, we show, for each H^2 category, coverage probabilities of both meta-analysis methods and proportions of simulation runs within each H^2 category. Information on the median I^2 value and overall coverage probability is printed on the top of each panel. Overall coverage is similar to previously published results [14]; generally, results of the Hartung–Knapp method are closer to the nominal coverage probability, whereas the DerSimonian–Laird method is conservative for small K if $\tau^2 = 0$ and anti-conservative if $\tau^2 > 0$. However, in very homogeneous simulation runs ($H^2 < 0.4$), we observe in most

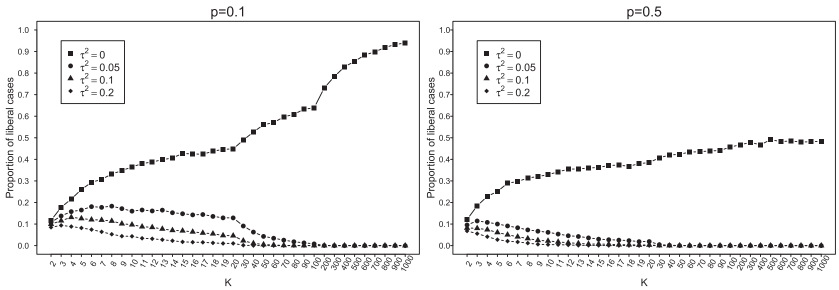


Figure 7. Proportion of simulations with shorter confidence interval for Hartung–Knapp method than fixed-effect meta-analysis

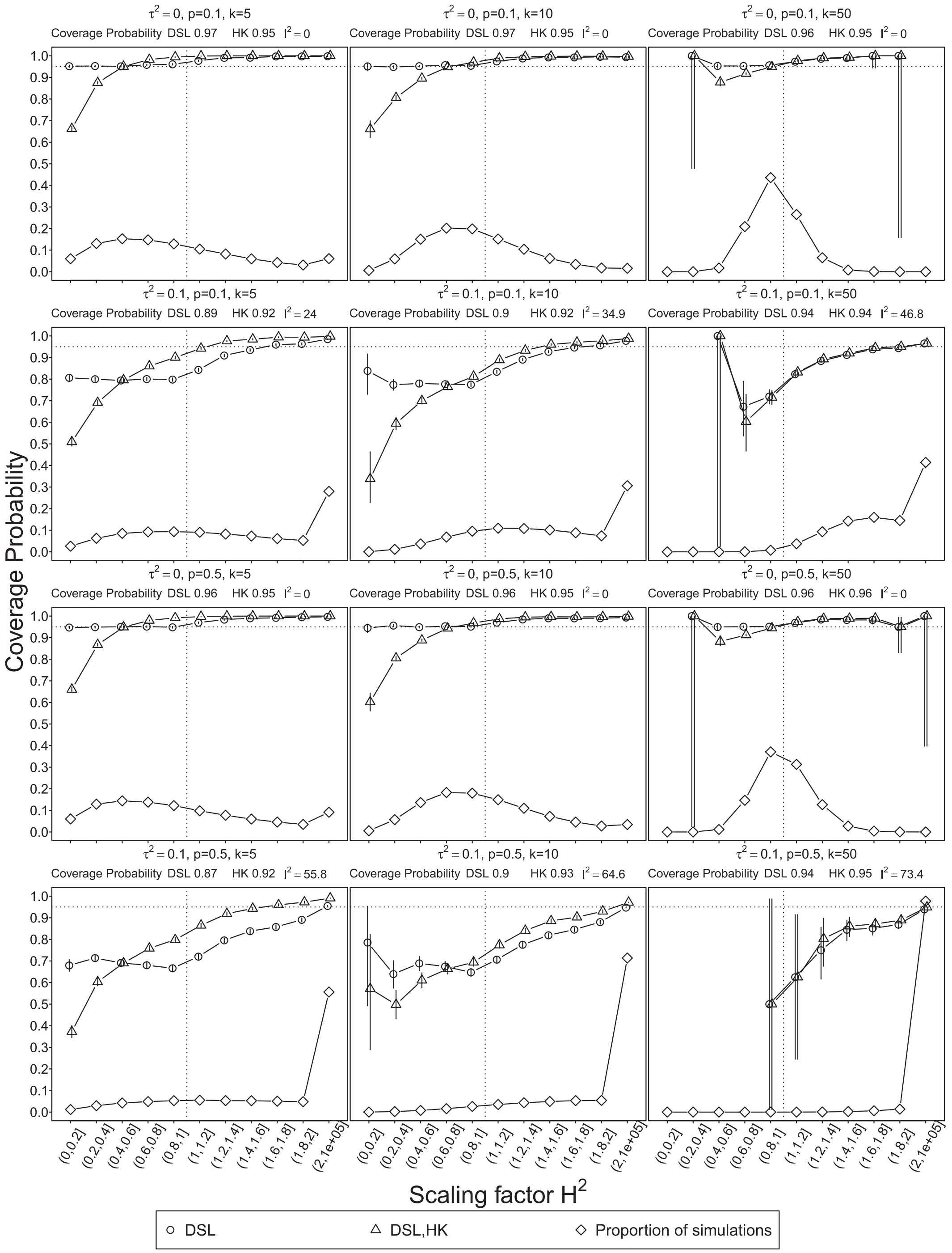


Figure 8. Comparison of the coverage probabilities of DerSimonian–Laird and Hartung–Knapp methods. The vertical lines represent the 95% confidence interval for the observed coverage probability. The I^2 values given represent the median of all values observed. DSL, DerSimonian–Laird; HK, Hartung–Knapp.

simulation scenarios that the coverage probability of the Hartung–Knapp method is smaller as compared with the classic random-effects meta-analysis. The proportion of very homogeneous simulation decreases with increasing K and τ^2.

Table I contains information on the ratio of confidence-interval length for fixed-effect meta-analysis and Hartung–Knapp method, in the second simulation setting. Values greater than one correspond to a shorter confidence interval for the Hartung–Knapp method. We calculated the proportion of runs above several threshold values: > 1, > 1.5, > 2, > 3, > 5. Note, the first column (> 1) contains the same information as Figure 7, albeit for the second simulation setting. The second column reports proportions

Table I. Proportion of simulations where the ratio of the lengths of fixed-effect and Hartung–Knapp confidence intervals is greater than the given value

			Ratio of the lengths of fixed-effect and Hartung–Knapp confidence intervals				
p	K	τ^2	> 1	> 1.5	> 2	> 3	> 5
0.1	5	0	0.26566	0.07167	0.02478	0.00578	0.00078
		0.1	0.12927	0.03138	0.01055	0.00219	0.00027
	10	0	0.36598	0.03613	0.00445	0.00021	0.00000
		0.1	0.09748	0.00587	0.00049	0.00002	0.00000
	50	0	0.55776	0.00010	0.00000	0.00000	0.00000
		0.1	0.00464	0.00000	0.00000	0.00000	0.00000
0.5	5	0	0.25966	0.07187	0.02619	0.00592	0.00084
		0.1	0.06069	0.01409	0.00482	0.00097	0.00014
	10	0	0.33471	0.03501	0.00431	0.00023	0.00001
		0.1	0.02109	0.00121	0.00010	0.00000	0.00000
	50	0	0.42919	0.00011	0.00000	0.00000	0.00000
		0.1	0.00002	0.00000	0.00000	0.00000	0.00000

where the fixed-effect confidence interval is at least 50% longer than the Hartung–Knapp confidence interval with largest values for $K = 5$ and $\tau^2 = 0$: 7.17% (for $p = 0.1$) and 7.19% ($p = 0.5$). For larger ratios, the proportion of runs with shorter Hartung–Knapp method confidence intervals is further reduced: below 3% (ratio > 2), below 1% (ratio > 3) and below 0.1% (ratio > 5). For ratios above 1.5, the largest proportions were always seen in simulation scenarios with $K = 5$ and $\tau^2 = 0$.

5. Discussion

In our empirical evaluation of the DerSimonian–Laird and Hartung–Knapp methods, we confirm the well-known property that the use of the Hartung–Knapp method results in most cases in a more conservative result as compared with the DerSimonian–Laird method with greater P-values for the test of an overall treatment effect and wider confidence intervals. In addition, we show in the empirical evaluation and simulations that the Hartung–Knapp method can result in much narrower confidence intervals and smaller P-values than the DerSimonian–Laird method in meta-analyses with very homogeneous study results. Actually, in these situations, the between-study variance is estimated to be zero, and thus, the comparison is between a fixed-effect model and the Hartung–Knapp random-effects model.

Looking at Eqs 10 and (8) to calculate the confidence interval for the Hartung–Knapp and DerSimonian–Laird methods, we observe that (i) the same treatment estimate is used in both equations and (ii) the quantile of the t-distribution is always greater or equal than the quantile of the standard normal distribution. Thus, a narrower confidence interval and correspondingly a smaller P-value can only result from a smaller standard error in the Hartung–Knapp method.

The two meta-analyses shown in Figures 2 and 3 are based on five and seven studies, respectively. The 97.5% quantile for a t-distribution with 4 and 6 degrees of freedom is 2.78 and 2.45, respectively. These values are about 40% and 25% greater than 1.96, the corresponding quantile of the standard normal distribution used in the calculation of the 95% confidence interval for the DerSimonian–Laird method. Nevertheless, the confidence interval of the Hartung–Knapp method is much narrower than the confidence interval of the fixed-effect model indicating a substantially smaller standard error for the Hartung–Knapp method.

As described before, if the estimated between-study variance is zero, we can show analytically that the standard error of the Hartung–Knapp method is always smaller or equal than the standard error of the DerSimonian–Laird method. If the between-study variance is greater than zero and it is assumed to be known, then the probability for a smaller standard error with the Hartung–Knapp method compared with the DerSimonian–Laird method is always greater than 0.5; this probability is approaching 0.5 as the number of studies in meta-analysis is increasing. Accordingly, in this situation, the estimation of the standard error with the Hartung–Knapp method is on average anti-conservative compared with the classic random-effects model. These analytical results support our view that the reason for the Hartung–Knapp method being conservative is mainly based on the use of a quantile of the t-distribution instead of the standard normal distribution. Concerns have been expressed [19] whether the Hartung–Knapp method is any better than simply replacing z-quantile with corresponding t-quantile in the classic random effects method. However, simulation studies showed [6, 7] that the Hartung–Knapp method holds the prespecified significance level well in all simulation scenarios, whereas the classic random effects method is conservative for few studies when the underlying between-study heterogeneity is low and gets liberal when heterogeneity increases. Therefore, replacing z-quantile with t-quantile in the classic random-effect method would yield too conservative results for low between-study heterogeneity.

In fact, the possibility that the standard error for the Hartung–Knapp method can be smaller than the standard error from the DerSimonian–Laird method has been noted more than 10 years ago by Knapp and Hartung in the context of random-effects meta-regression [14]. They proposed an *ad hoc* modification of the Hartung–Knapp variance estimate, which guarantees that the width of the confidence interval of the Hartung–Knapp method is always greater or equal than the confidence interval of the DerSimonian–Laird method. This modification has been implemented in the metareg command in Stata [20]; to our knowledge, this is the only implementation.

As we note in the description of the Hartung–Knapp method, an estimate of the between-study variance is needed to apply this method. For simplicity, we use the DerSimonian–Laird estimate of the between-study variance in this paper. In principle, any other estimation method for the between-study variance can be used to conduct the Hartung–Knapp method. We did not systematically evaluate the implication of using a different estimate of the between-study variance. As a sensitivity analysis, we ran simulations using the Paule–Mandel method [21], and results were similar to those presented for the DerSimonian–Laird method.

Based on our empirical evaluation and analytical results, we suggest that meta-analysts using the Hartung–Knapp method as the primary statistical approach should conduct a sensitivity analysis using the fixed-effect model. This sensitivity analysis can also be used as a first check whether small study effects are present in the meta-analysis, which are often an indication of publication bias; as the random-effects meta-analysis gives more weight to small studies than the fixed-effect model, a large different in point estimates between these two methods is an indication of small study effects [22–24].

The refined method has also been evaluated in the context of multivariate meta-analysis [17]. This paper also noted that the refined method may produce shorter confidence intervals than the fixed-effect model when the scaling factor H^2 is less than one. We did not cover multivariate meta-analysis in our paper, but the simulation study in [17] suggests that our results can be transferred to the multivariate setting.

In summary, we agree with previous statements [9, 11] that the Hartung–Knapp method appears superior to the use of a random-effects model based on the DerSimonian–Laird method. However, we suggest to use the fixed-effect model in a sensitivity analysis, which somewhat ironically is a special case of the DerSimonian–Laird method marked as inferior to the Hartung–Knapp method.

Acknowledgement

The authors thank Ekkehard Glimm, three anonymous reviewers and the associate editor for their helpful comments, which greatly improved this paper. The first author's work was supported by funding from the European Communitys Seventh Framework Programme FP7/2011: Marie Curie Initial Training Network, MEDIASRES (Novel Statistical Methodology for Diagnostic/Prognostic and Therapeutic Studies and Systematic Reviews). Gerta Rücker's work was supported by the Deutsche Forschungsgemeinschaft (grant number RU1747/1-1).

References

1. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
2. The Cochrane Collaboration. *Review Manager (RevMan) [Computer program]. Version 5.3*. Copenhagen: The Nordic Cochrane Centre, 2014.
3. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine* 2007; **26**: 37–52.
4. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuss O, Higgins JP, Langan D, Salanti G. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods* 2015. doi: 10.1002/jrsm.1164.
5. Hartung J. An alternative method for meta-analysis. *Biometrical Journal* 1999; **41**:901–916.
6. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine* 2001; **20**:1771–1782.
7. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine* 2001; **20**:3875–3889.
8. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine* 2002; **21**(21):3153–3159.

9. IntHout J, Ioannidis JPA, Borm GF. 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* 2014; **14**:25.
10. López-López JA, Botella J, Sánchez-Meca J, Marín-Martínez F. Alternatives for mixed-effects meta-regression models in the reliability generalization approach: A simulation study. *Journal of Educational and Behavioral Statistics* 2013; **38**(5):443–469.
11. Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR., Guallar E, Goodman SN. Random-effects meta-analysis of inconsistent effects: A time for change. *Annals of Internal Medicine* 2014; **160**(4):267–270.
12. Jüni P. Department of Social and Preventive Medicine, University of Berne, Switzerland. Personal communication, 2006.
13. Carpenter JR, Schwarzer G, Rücker G, Künstler R. Empirical evaluation showed that the Copas selection model provided a useful summary in 80% of meta-analyses. *Journal of Clinical Epidemiology* 2009; **62**:624–631.
14. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* 2003; **22**:2693–2710.
15. Schwarzer G. *meta: General package for meta-analysis*, 2015. http://CRAN.R-project.org/package=meta;https://github.com/guido-s/meta, R package version 4.1-0.
16. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 2010; **36**(3):1–48.
17. Jackson D, Riley RD. A refined method for multivariate meta-analysis and meta-regression. *Statistics in Medicine* 2014; **33**(4):541–554.
18. Boyle P, Flowerdew R, Williams A. Evaluating the goodness of fit in models of sparse medical data: a simulation approach. *International Journal of Epidemiology* 1997; **26**(3):651–656.
19. Copas J. Letters to the editor: a simple confidence interval for meta-analysis. *Statistics in Medicine* 2002; **21**:3153–3159.
20. Harbord RM, Higgins JPT. Meta-regression in Stata. *Stata Journal* 2008; **8**(4):493–519(27). http://www.stata-journal.com/article.html?article=sbe23_1.
21. Paule RC, Mandel J. Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards* 1982; **87**(5):377–385.
22. Poole C, Greenland S. Random-effects meta-analysis are not always conservative. *American Journal of Epidemiology* 1999; **150**(5):469–475.
23. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, Carpenter J, Rücker G, Harbord RM, Schmid CH, Tetzlaff J, Deeks JJ, Peters J, Macaskill P, Schwarzer G, Duval S, Altman DG, Moher D, Higgins JPT. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal* 2011; **343**:d4002. http://bmj.com/cgi/content/full/bmj.d4002, doi: 10.1136/bmj.d4002.
24. Rücker G, Carpenter J, Schwarzer G. Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal* 2011; **53**(2):351–368.

6 General discussion and conclusions

6.1 Arm based vs baseline contrast models and parametrizations

One of the main objectives of this thesis was to evaluate and extend the use of the arm-based modelling to different applications in medical research and drug development. As shown in the Introduction (Section 1.3.3), for the fixed-effects model the proposed modelling framework is the same but just using a different parametrization than the baseline contrast approach. For random-effects models the model itself is different leading to different assumption for the random effects' variance covariance matrix. Depending on the likelihood function and estimation method even in case of a random-effects model the two different assumptions can lead to exactly same or very similar point estimates of mean and variance parameters and yielding the same inference and conclusion of differences between the treatments of interest. This is discussed in more detail in the next section.

6.2 Estimation methods for between study variance

It has been noted that depending on the likelihood and link function the model with arm-based parametrization may lead (especially with maximum likelihood estimation) to under-estimation of between study variance, which leads to shorter confidence intervals for treatment contrast estimates and may lead to inflated type one error rate for treatment comparisons (Jones et al., 2011).

In case of a normal likelihood and identity link function this can be taken care of at the estimation stage by using restricted maximum likelihood which corrects for the loss of degrees of freedom and leads to similar result as baseline contrast model and gives unbiased estimates of the between study variance (Littell et al., 2006). For other types of responses (binomial and Poisson for example) restricted maximum likelihood is not available. Several techniques for estimating the between-study variance are available in common software packages like SAS and some of these can be used to ensure less biased between-study variance estimation as shown by simulation in (Piepho et al., 2018). For example, using the residual pseudo-likelihood estimation for binomial and count data is same type of estimation as REML for normally distributed response data. One should keep in mind that even though the residual pseudo-likelihood makes REML type of estimation available it is based on approximation and has limitations if the binomial sample sizes are small. In practice this is often not a problem in clinical trials as the sample sizes are relatively large, especially for late stage confirmatory clinical trials. However, even in network meta-analysis of large phase 3 trials it might become a problem if the event itself is rare, like for example some rare adverse event which occurs less than 1 in 1000. The challenges of rare events and how to overcome it has been discussed in several recent published research papers (Efthimiou et al., 2019; Günhan et al., 2020; Zabriskie et al., 2021).

Regardless of baseline contrast model or arm based model or parametrization the estimation of between study variance in meta-analysis has been subject to extensive research (Veroniki et al., 2016). As the number of studies in meta-analysis is often limited, the between-study variance estimation has limitations. For example, if a simple two-treatment meta-analysis contains 3 studies, the degrees of freedom for between study variance estimation is 2, leading to a very wide distribution of the between-study variance estimate and hence considerable uncertainty about the heterogeneity between studies.

Chapter 5 was based on evaluation different between study variance methods in traditional two-treatment meta-analysis. The main finding of the paper was that in some cases random effect meta-analysis using Hartung-Knapp method may yield shorter confidence intervals for combined treatment effect than fixed effect meta-analysis. This was likely to happen if the observed differences between average treatment effect estimates was less than one would expect based on the within study variance of included trials.

6.3 Model selection in non-proportional hazards NMA

The second real life example, presented in Chapter 4, contained an example of using the framework for a more complicated modelling problem. This work was motivated by the need for more efficient ways to apply the method of fractional polynomials NMA (Jansen, 2011). The method by Jansen is often used for evidence synthesis of immunotherapies (Herbst et al., 2021; Schulz et al., 2019) and to support decision making at different health technology assessment agencies like NICE (The National Institute for Health and Care Excellence). For this type of model it has proven to be difficult to find starting values for parameters. Furthermore, exploration of model convergence can be time consuming. The model building approach presented in Chapter 4 has been applied for example in comparison of different treatments for HER2-positive metastatic breast cancer after HER2-targeted therapy (DeBusk et al., 2021). It has been also acknowledged in (Freeman et al., 2022) that the proposed method has potential to speed up the model selection process, however they address the need for further evaluation of similarity of using the Akaike's Information Criterion (AIC) versus the deviance information criterion (DIC). Due to the analogy of AIC and DIC this is likely not to be an issue in case of fixed effect models with uninformative priors. Generally, the model selection criteria and algorithms for these types of model would require further research. For example, it is not guaranteed that same fractional polynomial or piecewise exponential model would be selected within fixed or random effects NMA. The challenges for selecting the best model to take into account the non-proportionality of the hazards apply to both frequentist and Bayesian models. As many models may need to be explored, the frequentist ANOVA type of model enables rapid exploration of different models and is also less prone to human errors in coding of the complicated models.

6.4 Level of aggregation

The manuscript in Chapter 2 explored the network meta-analysis in a setting where some of the trials provide individual patient data and some only aggregated data. In this setting the aggregated data may be often sufficient if individual trials are analysed in a consistent manner and variance estimates of treatments means are given. In principle, if the aggregated trials report sufficient statistics for the NMA the results of the NMA will be equal from both individual patient data NMA and aggregated data NMA. Similar comparisons of single-stage (individual patient data meta-analysis) and two-stage (aggregated study level data meta-analysis) has also been discussed in the context of multi-environment trials with similar conclusions (Damesa et al., 2017). If some treatment effect modifying covariates are present and needed in analysis, then having individual patients data is often necessary to be able to reliably estimate treatment effects at different covariate values.

In the third application presented in Chapter 4 it is possible to extract the individual patient data for the response variable by using digitalization of published Kaplan-Maier curves and algorithms to recover the individual survival times (Guyot et al., 2012). Even in this case the challenge of taking into account the impact of potential differences in baseline covariates remains.

In case of having access to individual patient data for one of the studies and aggregated data for other studies, like in the example of Chapter 2 it is possible to weigh the individual patient data to “match” the aggregated baseline covariates using for example so called matching adjusted in direct comparison(MAIC) method (Signorovitch et al., 2012). For meta-analysis and network-meta analysis with more than two studies the challenge of having several different baseline populations remains. Also, all relevant baseline covariates may not be collected or reported in a similar way in different studies.

6.5 Software implementations

For frequentist analysis of GLM with arm-based parametrization basically any standard software package can be used. This thesis used both SAS and R for fitting the models. Both software have their advantages. In drug development SAS has traditionally been the software used for reporting clinical trials, however R enjoys increasing popularity and is also accepted by health authorities.

The first real life application of the network meta-analysis presented in Chapter 3 included over 200 trials and 60 treatments, with ANOVA type of model with arm-based parametrization and frequentist estimation the model was possible to be fitted with standard SAS procedures with minimal coding and computational time. Specifying this extensive application with baseline contrast model and Bayesian estimation method would have required priors and starting values for 200 trials and 60 treatments leading to very extensive computation.

Of note, there is no reason why the ANOVA based methods for meta-analysis and network meta-analysis couldn't be fitted in a Bayesian framework as well. This can be done for example using the SAS procedure PROC BGLIMM (Piepho and Madden, 2022; Rott et al., 2021).

6.6 Future perspectives

One of the most topical research interests in drug development is how to use external control arms to provide evidence on the efficacy of new drugs compared to existing treatment options (Burger et al., 2021; Davi et al., 2020; Schmidli et al., 2020). The development of use of external control data has many issues in common with indirect treatment comparisons, one of the most important being that the treatments are compared in similar populations and the differences in potential confounding patient characteristics and different clinical practices are properly considered in analyses. Also, the question of how to include trials with outrandomized control arm into existing networks could be a potential topic for future research.

Traditionally the aim of statistics in medical publications has been to aggregate the collected data for making more general conclusions of the research question. However, usually when data is aggregated, some information is also lost. Given the development of new digital tools for exploring information from different sources, it has become possible to extract more precise data from figures, which are showing some aggregated data, for example. Building on methods to efficiently extract the data given in summary figures could open possibilities to make more precise analyses and evidence synthesis of older publications.

Giving the increasing number of treatment options for many diseases it is important, when developing a new drug for a disease, to be able to compare in objective way the existing treatments and show some benefit compared to existing treatment options. Combining evidence synthesis methods efficiently with adaptive designs could enable making better informed decision whether drug development programs of new compounds should be continued or discontinued or if there is need for adjustment in development programs in order to fill the most important unmet clinical needs with the newly developed drugs.

6.7 Conclusions

This thesis explored and developed the use of generalized linear mixed models in a setting of network meta-analysis of randomized clinical trials. In practice the most popular analysis method in the field of network meta-analysis has been the baseline contrast model which is usually fitted in a Bayesian framework. The baseline contrast model and Bayesian estimation provides great flexibility, but also comes with some unnecessary complications for certain types of analyses.

This thesis showed how methods originally developed and extensively used in agricultural research can be used in network meta-analysis of clinical trials providing efficient calculation, estimation, and inference. Some of the examples used in this thesis arose from analyses needed for real applications in drug development and were directly used in medical research.

7 Summary

Network meta-analyses of published clinical trials has received increased attention over the past years with some meta-analytic publications having had a big impact on the cost-benefit assessment of important drugs. Much of the research has been based on Bayesian analysis using so called base-line contrast model. The research in network meta-analysis methodology has in parts been isolated from other fields of mathematical statistics and is lacking an integrative framework clearly separating statistical models and assumptions, inferential principles, and computational algorithms. The very extensive past research on ANOVA and MANOVA of un- balanced designs, variance component models, generalised linear models with fixed and/or random effects, provides a wealth of useful approaches and insights. These models are especially common in agricultural statistics and this thesis extended the use of the general statistical methods mainly applied in agricultural statistics to applications of network meta-analysis of clinical trials.

The methods were applied to four different research problems in separate manuscripts.

The first manuscript was based on a simulated case (based on real example) where some of the trials provided individual patient data and some only aggregated data. The outcome type considered was continuous normally distributed data. This manuscript provides models for jointly model the individual patient data and aggregated data. It was also explored how much information is lost if data is aggregated and how to quantify the amount of lost information.

The second manuscript was based a real life dataset with pain medications used in acute postoperative pain. The outcome of interest was binomial, whether a subject experienced pain relief or not. The dataset used for NMA included 261 trials with 52 different treatment and dose combinations, making it extraordinarily rich and large network.

The third manuscript developed methods for a case of time-to-event-outcome extracted from published Kaplan-Meier curves of survival analyses. This re-generated individual patient data was then used to model and compare the Kaplan-Meier curves and hazards of different treatments.

The fourth manuscript of the thesis was tackling the problem of between-trial variance estimation for a specific method of Hartung-Knapp in classical two-treatment meta-analysis. The main finding of the paper was that in some cases random effect meta-analysis using Hartung-Knapp method may yield shorter confidence intervals for combined treatment effect than fixed effect meta-analysis and therefore the recommendation is to always compare results from Hartung-Knapp method with fixed effect meta-analysis.

This thesis explored and developed the use of generalized linear mixed models in a setting of network meta-analysis of randomized clinical trials. In practice the most popular analysis method in the field of network meta-analysis has been the baseline contrast model which is usually fitted in a Bayesian framework. The baseline contrast model and Bayesian estimation provides great flexibility, but also come with some unnecessary complications for certain types of analyses.

This thesis showed how methods originally developed and extensively used in agricultural research can be used in other field providing efficient calculation, estimation, and inference. Some of the examples used in this thesis arose from analyses needed for real applications in drug development and were directly used in medical research.

8 Zusammenfassung

In den letzten Jahren haben Netzwerk-Meta-Analysen von publizierten Ergebnissen klinischer Studien viel Aufmerksamkeit erhalten und die Kosten-Nutzen-Einschätzung wichtiger pharmazeutischer Präparate in erheblichem Umfang beeinflusst. Ein Großteil der methodischen Forschung zur Meta-Analyse konzentrierte sich dabei auf Bayessche Methoden im sogenannten Baseline-Contrast-Modell. Diese methodischen Untersuchungen haben z.T. losgelöst von anderen Bereichen der mathematischen Statistik stattgefunden. Daher fehlte ein integrativer Rahmen, welcher mathematische Modelle und Annahmen, Prinzipien der Inferenz und Algorithmen zur Ermittlung von Effektschätzungen klar voneinander trennte. Die sehr umfangreichen Erkenntnisse zur Varianzanalyse (ANOVA und MANOVA) unbalanzierter Versuchsanordnungen, Varianzkomponentenmodellen sowie generalisierten linearen Modellen mit festen und zufälligen Effekten, welche in der Vergangenheit, nicht zuletzt im Bereich der Agrarwissenschaften, erlangt wurden, sind auch für die Methodik der Meta-Analyse sehr nützlich. Diese Arbeit erweitert die Nutzung solcher Methoden auf die Netzwerk-Meta-Analyse klinischer Studien.

Die Anwendung dieser Methoden wird in vier Manuskripten dieser kumulativen Thesis dargestellt.

Im ersten Manuskript wird eine Situation untersucht, bei der für einen Teil der untersuchten klinischen Studien individuelle Patientendaten (IPD) vorliegen, für einen anderen Teil indes nur aggregierte Daten (AD). Das Manuskript stellt Modelle vor, welche sich für die gemeinsame Analyse solcher Daten eignen. Es wird angenommen, dass die Daten Normalverteilungen entstammen. Die Daten wurden basierend auf realen Studiendaten simuliert. Das Manuskript untersucht, wieviel Information durch die Datenaggregation verloren geht und wie dieser Informationsverlust quantifiziert werden kann.

Das zweite Manuskript untersucht einen Datensatz aus 261 klinischen Studien, in denen insgesamt 52 verschiedene Behandlungen gegen akute postoperative Schmerzen geprüft wurden. Die Zielgröße ist binär und hält fest, ob Schmerzlinderung erzielt wurde oder nicht. Aufgrund der vielen Studien und Behandlungen liegt hier ein aussergewöhnlich umfangreiches und komplexes Netzwerk vor.

Im dritten Manuskript werden Methoden zur Analyse von Überlebenszeitdaten vorgestellt. Die Daten wurden mithilfe von Softwaretools aus publizierten Kaplan-Meier-Kurven extrahiert. Die so gewonnenen individuellen Patientendaten wurden benutzt, um die

Überlebenskurven zu modellieren und die Hazardraten verschiedener Behandlungen zu vergleichen.

Das vierte Manuskript betrachtet einen speziellen Aspekt der Inter-Studien-Varianzschätzung in der klassischen Meta-Analyse mit zwei Behandlungsarmen. Das Hauptergebnis dieser Untersuchung ist, dass die sogenannte Hartung-Knapp-Methode in Modellen mit zufälligen Effekten in bestimmten Fällen zu kürzeren Konfidenzintervallen für die kombinierte Behandlungseffektschätzung führen kann als die entsprechende Schätzung in einem Modell mit festen Effekten. Daher wird empfohlen, in konkreten Analysen beide Methoden zu verwenden und die Ergebnisse zu vergleichen.

Übergreifendes Thema dieser Thesis ist die Untersuchung generalisierter linearer gemischter Modelle für Netzwerk-Meta-Analysen klinischer Studien. In der Praxis ist in diesem Bereich das Baseline-Kontrast-Modell mit Bayesschen Effektschätzungen das populärste Modell. Dieses Modell und die Methode der Bayes-Schätzung erlauben hohe Flexibilität, aber in manchen Fällen verkomplizieren sie die Analyse auf unnötige Weise.

Diese Arbeit zeigt, wie Methoden, die ursprünglich in den Agrarwissenschaften entwickelt wurden und ausgiebig genutzt werden, auch für die Meta-Analyse klinischer Studien effiziente Schätz- und Inferenzmethoden zur Verfügung stellen. Einige der Beispiele in dieser Arbeit sind durch Anwendungen in der Medikamentenentwicklung motiviert und wurden bereits in konkreten medizinischen Forschungsvorhaben eingesetzt.

9 References

Burger, H.U., Gerlinger, C., Harbron, C., Koch, A., Posch, M., Rochon, J., Schiel, A., 2021. The use of external controls: To what extent can it currently be recommended? Pharm. Stat. 20, 1002–1016. https://doi.org/10.1002/pst.2120

Curran, P.J., Hussong, A.M., 2009. Integrative data analysis: The simultaneous analysis of multiple data sets. Psychol. Methods 14, 81–100. https://doi.org/10.1037/a0015914

Damesa, T.M., Möhring, J., Worku, M., Piepho, H.-P., 2017. One Step at a Time: Stage-Wise Analysis of a Series of Experiments. Agron. J. 109, 845–857. https://doi.org/10.2134/agronj2016.07.0395

Davi, R., Mahendraratnam, N., Chatterjee, A., Dawson, C.J., Sherman, R., 2020. Informing single-arm clinical trials with external controls. Nat. Rev. Drug Discov. 19, 821–822. https://doi.org/10.1038/d41573-020-00146-5

DeBusk, K., Abeysinghe, S., Vickers, A., Nangia, A., Bell, J., Ike, C., Forero-Torres, A., Blahna, M.T., 2021. Efficacy of tucatinib for HER2-positive metastatic breast cancer after HER2-targeted therapy: a network meta-analysis. Future Oncol. https://doi.org/10.2217/fon-2021-0742

Efthimiou, O., Rücker, G., Schwarzer, G., Higgins, J.P.T., Egger, M., Salanti, G., 2019. Network meta-analysis of rare events using the Mantel-Haenszel method. Stat. Med. 38, 2992–3012. https://doi.org/10.1002/sim.8158

Freeman, S.C., Cooper, N.J., Sutton, A.J., Crowther, M.J., Carpenter, J.R., Hawkins, N., 2022. Challenges of modelling approaches for network meta-analysis of time-to-event outcomes in the presence of non-proportional hazards to aid decision making: Application to a melanoma network. Stat. Methods Med. Res. 09622802211070253. https://doi.org/10.1177/09622802211070253

Günhan, B.K., Röver, C., Friede, T., 2020. Random-effects meta-analysis of few studies involving rare events. Res. Synth. Methods 11, 74–90. https://doi.org/10.1002/jrsm.1370

Guyot, P., Ades, A., Ouwens, M.J., Welton, N.J., 2012. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med. Res. Methodol. 12, 9. https://doi.org/10.1186/1471-2288-12-9

Hasselblad, V., 1998. Meta-analysis of Multitreatment Studies. Med. Decis. Making 18, 37–43. https://doi.org/10.1177/0272989X9801800110

Herbst, R., Jassem, J., Abogunrin, S., James, D., McCool, R., Belleli, R., Giaccone, G., De Marinis, F., 2021. A Network Meta-Analysis of Cancer Immunotherapies Versus Chemotherapy for First-Line Treatment of Patients With Non-Small Cell Lung Cancer and High Programmed Death-Ligand 1 Expression. Front. Oncol. 11, 676732. https://doi.org/10.3389/fonc.2021.676732

Jansen, J.P., 2011. Network meta-analysis of survival data with fractional polynomials. BMC Med. Res. Methodol. 11, 61. https://doi.org/10.1186/1471-2288-11-61

Jones, B., Roger, J., Lane, P.W., Lawton, A., Fletcher, C., Cappelleri, J.C., Tate, H., Moneuse, P., 2011. Statistical approaches for conducting network meta-analysis in drug development. Pharm. Stat. 10, 523–531.

Koricheva, J., Gurevitch, J., Mengersen, K., 2013. Handbook of meta-analysis in ecology and evolution. Princeton University Press.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., Oliver, S., 2006. SAS for mixed models. SAS publishing.

Lu, G., Ades, A., 2006. Assessing evidence inconsistency in mixed treatment comparisons. J. Am. Stat. Assoc. 101.

Madden, L.V., Piepho, H.-P., Paul, P.A., 2016. Statistical Models and Methods for Network Meta-Analysis. Phytopathology® 106, 792–806. https://doi.org/10.1094/PHYTO-12-15-0342-RVW

Piepho, H.-P., Madden, L.V., 2022. How to observe the principle of concurrent control in an arm-based meta-analysis using SAS procedures GLIMMIX and BGLIMM. Res. Synth. Methods 13, 821–828. https://doi.org/10.1002/jrsm.1576

Piepho, H.-P., Madden, L.V., Roger, J., Payne, R., Williams, E.R., 2018. Estimating the variance for heterogeneity in arm-based network meta-analysis. Pharm. Stat. 17, 264–277. https://doi.org/10.1002/pst.1857

Piepho, H.P., Williams, E.R., Madden, L.V., 2012. The Use of Two-Way Linear Mixed Models in Multitreatment Meta-Analysis. Biometrics 68, 1269–1277. https://doi.org/10.1111/j.1541-0420.2012.01786.x

Rott, K.W., Lin, L., Hodges, J.S., Siegel, L., Shi, A., Chen, Y., Chu, H., 2021. Bayesian meta-analysis using SAS PROC BGLIMM. Res. Synth. Methods 12, 692–700. https://doi.org/10.1002/jrsm.1513

Schmidli, H., Häring, D.A., Thomas, M., Cassidy, A., Weber, S., Bretz, F., 2020. Beyond Randomized Clinical Trials: Use of External Controls. Clin. Pharmacol. Ther. 107, 806–816. https://doi.org/10.1002/cpt.1723

Schulz, C., Gandara, D., Berardo, C.G., Rosenthal, R., Foo, J., Morel, C., Ballinger, M., Watkins, C., Chu, P., 2019. Comparative Efficacy of Second- and Subsequent-line Treatments for Metastatic NSCLC: A Fractional Polynomials Network Meta-analysis of Cancer Immunotherapies. Clin. Lung Cancer 20, 451-460.e5. https://doi.org/10.1016/j.cllc.2019.06.017

Senn, S., 2000. The many modes of meta. Drug Inf. J. 34, 535–549.

Signorovitch, J.E., Sikirica, V., Erder, M.H., Xie, J., Lu, M., Hodgkins, P.S., Betts, K.A., Wu, E.Q., 2012. Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research. Value Health 15, 940–947. https://doi.org/10.1016/j.jval.2012.05.004

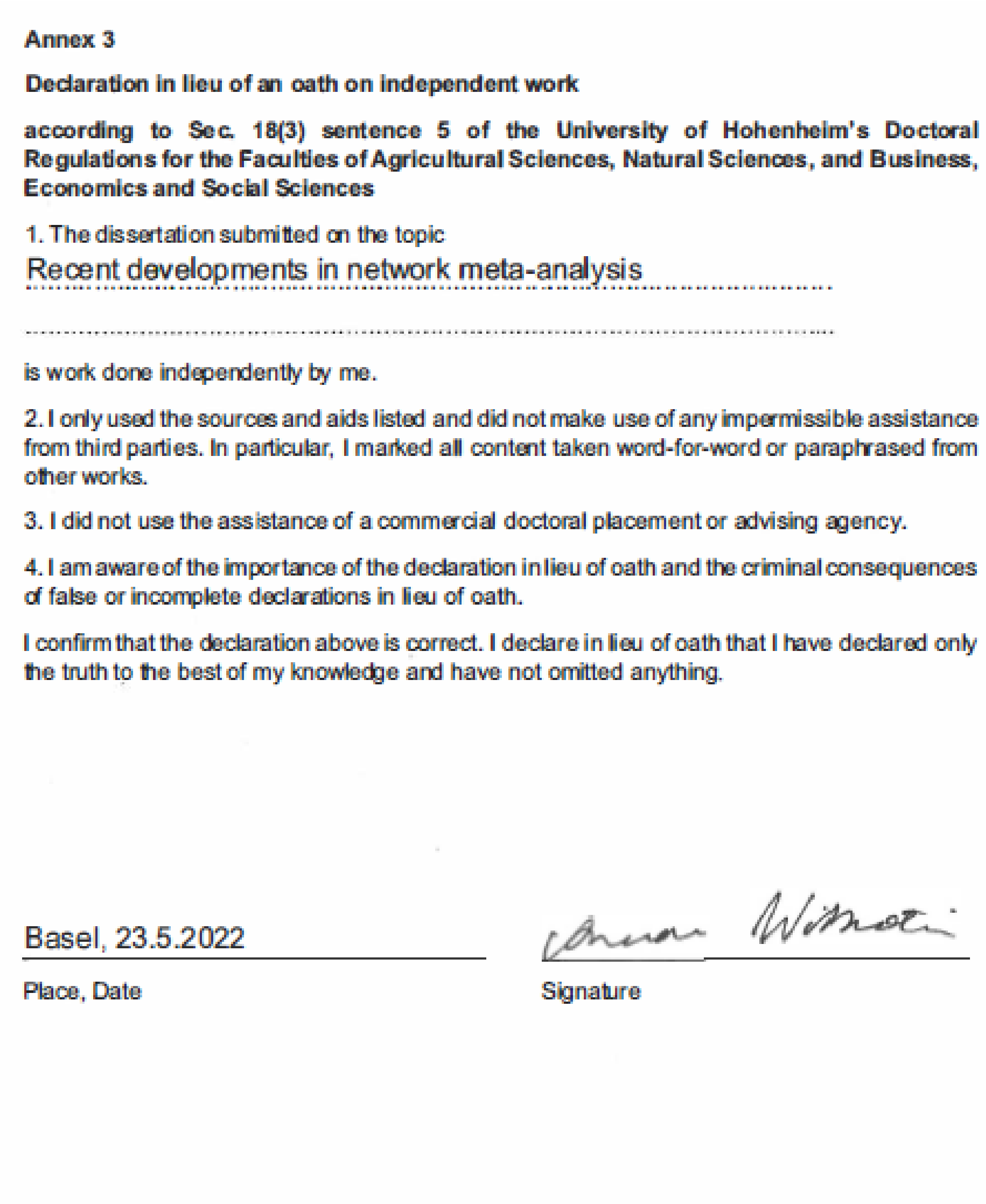
Veroniki, A.A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J.P., Langan, D., Salanti, G., 2016. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res. Synth. Methods 7, 55–79. https://doi.org/10.1002/jrsm.1164

Wiksten, A., Hawkins, N., Piepho, H.-P., Gsteiger, S., 2020. Nonproportional Hazards in Network Meta-Analysis: Efficient Strategies for Model Building and Analysis. Value Health 23, 918–927. https://doi.org/10.1016/j.jval.2020.03.010

Yates, F., Cochran, W.G., 1938. The analysis of groups of experiments. J. Agric. Sci. 28, 556–580. https://doi.org/10.1017/S0021859600050978

Zabriskie, B.N., Corcoran, C., Senchaudhuri, P., 2021. A comparison of confidence distribution approaches for rare event meta-analysis. Stat. Med. 40, 5276–5297. https://doi.org/10.1002/sim.9125

Affidavit



Annex 3

Declaration in lieu of an oath on independent work

according to Sec. 18(3) sentence 5 of the University of Hohenheim's Doctoral Regulations for the Faculties of Agricultural Sciences, Natural Sciences, and Business, Economics and Social Sciences

1. The dissertation submitted on the topic

Recent developments in network meta-analysis

is work done independently by me.

2. I only used the sources and aids listed and did not make use of any impermissible assistance from third parties. In particular, I marked all content taken word-for-word or paraphrased from other works.

3. I did not use the assistance of a commercial doctoral placement or advising agency.

4. I am aware of the importance of the declaration in lieu of oath and the criminal consequences of false or incomplete declarations in lieu of oath.

I confirm that the declaration above is correct. I declare in lieu of oath that I have declared only the truth to the best of my knowledge and have not omitted anything.

Basel, 23.5.2022

Place, Date

Signature