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

Ingeborg van der Tweel*, Konstantinos Pateras, G. Caroline M. van Baal, Kit C.B. Roes

Department of Biostatistics and Research Support, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

***Corresponding author:** Ingeborg van der Tweel, e-mail: i.vandertweel@umcutrecht.nl

Abstract

Background: A randomized controlled trial is considered the gold standard in clinical research, also in rare diseases. However, in small populations, single large-scale well-powered trials are often not possible. For valid decision-making, both on efficacy and on safety, evidence generated from series of trials could be exploited.

Methods: We conducted a review over a five-year period between January 2009 and December 2013 to identify relevant methodology on combining results of series of trials.

Results: Sixty-two papers were identified and summarized in this review. Its focus is on frequentist methodology. Most papers deal with meta-analyses on aggregated data. Only few papers discuss multivariate outcomes. We categorized the relevant methods according to the type of (meta-) analysis.

Conclusions: Only a few papers dealt directly with series of trials in small populations. The results of the review lead to some directions for further investigation on evidence-based decision-making from a (small) number of trials in small populations.

Keywords: clinical trials, frequentist methodology, meta-analysis, small populations

Background

Recently, the European Union funded the Asterix project: **A**dvances in **S**mall **T**rials d**E**sign for **R**egulatory **I**nnovation and e**X**cellence [OR1-2] to develop and implement innovative statistical methodologies for the evaluation of orphan drug treatments with clinical trials. Please note that the list of references is split into two parts: 1) references to papers in the review (RR) and 2) references to other publications (OR).

Rare diseases influence only a small part of the human population with a prevalence below 5 per 10,000 people in the European Community [OR3]. To evaluate the effect of a (new) intervention, a randomized controlled trial (RCT) is considered the gold standard also in rare diseases. However, large-scale well-powered RCTs are often not possible. To obtain sufficient, valid evidence for decision-making, both on efficacy and on safety, alternative methodological approaches have to be sought. Existing guidance [OR4-6] discusses and recommends designs that are suitable for single trials in small populations. To cope with the problem of small numbers of patients available for a single trial, evidence generated from series of trials in small populations could be exploited.

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

Discussion section, we will assess the usefulness and limitations of the described methods in a small number of small clinical trials (SCT).

Methods

We conducted a review to identify relevant methodology on combining results of series of trials as published between 1-1-2009 and 31-12-2013. Eligible studies were identified with several search strategies. First, we created a list of landmark papers, i.e. specific papers we wanted to be found and included in our final set of papers. Then we created a search strategy for PubMed. Because of the methodological nature of our review, clearly papers were missed by searching PubMed alone. We then extended our search to Web of Science (Science Citation Index (SCI)), Scopus, JSTOR and, lastly, the Cochrane Library (see Appendix for the search strategies). The search resulted in 8183 papers, of which 1031 from PubMed, 2438 from Scopus, 2230 from Web of Science, 2436 from JSTOR and 48 from Cochrane (considering only methodological articles) (see Figure 1 for the flow diagram of the search strategies). First of all, duplicates were removed. Then papers were excluded by journal, title and abstract based on their methodological relevance for the review. Two reviewers (KP, IvdT) independently scored the 358 articles from the remaining studies by judging title and abstract to include (I), probably include (PI), probably not include (PNI) and to exclude (E). Articles with a concordant score from both reviewers were either included (I or PI) or excluded (PNI or E) from the pool. The discordant 52 ones were discussed with a third independent colleague (GCMvB). From these 52, 38 were excluded and 14 were included into the final set of articles. Some papers with no abstract were considered for full reading; when they consisted of letters to the editor or commentaries on papers not included in the review they were excluded. One review article to which we could not get access was excluded. We only included papers written in English. Excluded were online abstracts only, books or book chapters, papers describing applications of meta-analyses and papers on n-of-1 trials.

To see whether we missed some general papers, we performed a search on the keyword "meta-analysis", restricting the search to a "trial". This resulted, however, in far too many papers to be discussed. We then specifically searched for papers on meta-analysis of small studies.

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

Results

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

General meta-analysis (MA)

A meta-analysis (MA) of RCTs is a statistical method to pool the results of several individual trials in a certain disease and summarize them into a point and its confidence interval (CI) estimate. The corresponding model for pooling k trials can be written as: $Y_i = \vartheta_i + \varepsilon_i$ with $\vartheta_i = \vartheta + \delta_i$ for i = 1, …, k.

Here, Y_i measures the treatment effect for trial i, ϑ_i is the trial-specific effect size, $\varepsilon_i \sim N(0; \sigma_i^2)$ with σ_i^2 the variance within the ith trial and $\delta_i \sim N(0; \tau^2)$ with τ^2 the between-trial variance.

For the estimates, a fixed effect (FE) and a random effects (RE) approach are distinguished. A FE model assumes that the unknown parameter value ϑ is the same for all trials; a RE model assumes that parameter values ϑ_i for the pooled trials follow some distribution. Both the within-trial variance σ_i^2 and the between-trial variance or heterogeneity parameter τ^2 have to be estimated from the trials.

Nowadays, MA methodology is widely implemented [OR7]. An important issue for the reliability of the results of an MA is the similarity of patients and other trial characteristics across the pooled set of RCTs. Trials can differ in patient-level and in study-level variables. Aiello et al [RR1] present graphical and analytical tools to identify quantitative criteria to detect these covariate imbalances. These tools are, however, not feasible in an MA with only few trials. Verbeek et al [RR2] stress that the credibility of an MA depends on the conceptual similarity of the studies and on the statistical heterogeneity.

An MA requires an extensive and complete systematic review (SR) of the medical literature on the disease concerned. RCTs with ‘negative’ results, i.e. no significant difference between the treatments compared, are less likely to be published, thus leading to publication bias. Publication bias can lead to ‘biased’ priors for both frequentist and Bayesian analyses and financial disclosure should be a covariate in meta-analyses to prevent investigator bias and assess uncertainty about study effects [RR3].

MA using a FE approach can lead to substantial inflation of the type I errors [RR4,OR8]. Both Higgins et al [RR5] and Chung et al [RR6] advise against the use of a FE or common model, but also against the testing of homogeneity. They emphasize that the naive presentation of only the mean μ of the RE analysis is misleading and estimation of the between-trial variance is just as important as well as its incorporation in a CI for μ. Both for frequentist and Bayesian inference from, let’s say, k trials, Higgins et al [RR5] (as well as Borenstein et al [OR7]) propose the use of a prediction interval based on a *t*-distribution with k-2 degrees of freedom instead of a Normal distribution to account for the uncertainty in the estimated τ^2. Chung et al [RR6] discuss the estimation of the between-study variance for small numbers of studies. Commonly used estimators then frequently result in a value of 0, thereby underestimating the true heterogeneity. They, following Borenstein et al [OR7] and Higgins et al [RR5], prefer a Bayesian informative prior distribution for the between-study variance based on plausible values from other, similar MAs or on historical data. They propose a Bayes modal estimator and compare its properties to those of other estimators. When study-level covariates are available, meta-regression analysis can be applied to decrease the heterogeneity.

RE models have disadvantages and may add unnecessary complexity to the analysis. To judge whether an RE or linear mixed effects model is appropriate, Demidenko et al [RR7] propose the RE coefficient of determination, the proportion of conditional variance explained by the heterogeneity of the studies in an MA.

For normally distributed outcomes standard MA theory assumes that variances are known. This theory is often applied to effect sizes with skewed distributions with variances to be estimated. Malloy et al [RR8**]** suggest to first apply a variance stabilizing transformation and then estimate point and interval parameters of FE or RE models using stable weights or profile approximate likelihood intervals. Further, a simple *t*-interval provides very good coverage of an overall effect size without estimation of the heterogeneity.

Viechtbauer [RR9] provides an extensive overview of the capabilities of the ‘metafor’ package for conducting meta-analyses with R.

Design

Journal guidelines state that a report of an RCT should include a summary of previous research findings, preferable an SR and MA, and explain how the new trial affects this summary. Such a summary should inform critical design issues such as sample size determination [RR10].
Sutton et al [RR10] stress that the existing evidence-base should be analysed in a more detailed way, e.g. by including individual patient data (IPD) in a MA, to be able to design future research more efficiently. The contribution of a newly planned RCT to the total evidence is evaluated through its incorporation into an updated MA in various ways. A new trial can be designed and powered in isolation based on the results of a MA or based on the statistical significance of the updated MA. Heterogeneity between RCTs can seriously influence the power of the updated MA. To better estimate heterogeneity, multiple small new studies can be preferred to a single large study containing the same number of subjects. This is an important issue in the design of new RCTs in rare diseases.
Goudie et al [RR11] found that only few published RCTs reported the use of previous trials to design a future trial and estimate its sample size. They also highlight the importance of adequately considering heterogeneity among studies in an MA, but note that between-study heterogeneity will often be estimated with poor precision. They point out that the process of using evidence from related, but not identical, studies could be formalized by more sophisticated modelling, such as the use of mixed treatment comparison (MTC) MA or of patient-level covariates. Ioannidis and Karassa [RR12] also emphasize the need to consider breadth, timing and depth of all evidence, including unpublished and on-going studies, for an SR and MA. They consider results from single, early stopped trials unreliable because of chance findings due to multiple testing and inflated treatment effect estimates.
Rotondi and Donner [RR13] describe estimation of an appropriate sample size for a planned cluster randomized trial by considering the role of the planned trial in a future MA. Sample size estimation can be based on power or reduction in variance or the perspective of non-inferiority. Their approach is based on simulated data using prior distributions for the intra-cluster correlation coefficient, the cluster size and the control event rate. An FE model with dichotomous outcomes is assumed as well as the availability of IPD. They suggest that their method ‘may prove particularly useful when dealing with a meta-analysis of a small number of studies’.

Heterogeneity

Between-trial variability or heterogeneity can be tested and estimated. A commonly used measure for heterogeneity between trials pooled in an MA is I^2. The statistic I^2 is the ratio of true heterogeneity τ^2 to the total variance ($\tau^2 + \sigma^2$). Higgins and Thompson [OR9] derived this measure assuming that all within-trial variances σ_i^2 were equal to σ^2, thus giving all trials the same weight, an assumption that is not met in most MAs. As a better alternative, Wetterslev et al [RR14] propose a measure of diversity D^2 to describe the relative variance reduction when the MA model changes from an RE MA to an FE MA. They show that $D^2 \geq I^2$ and thus, in general, this will lead to a larger information size, i.e. the required number of participants in an MA. The derivation of D^2, however, assumes that the FE population average is equal to the RE population average, which requires additional information if this assumption is not met.

Standard meta-analysis methods ignore the uncertainty in the estimation of the heterogeneity parameter [RR15,RR6]. Chung et al [RR6] describe the use of a profile likelihood function (following [OR10]) to construct a CI for μ or a Wald-type interval based on the observed instead of the expected information.

Turner et al [RR16] present a method to adjust for differences in rigour (i.e. lack of internal bias) and relevance (i.e. lack of external bias) between studies pooled in an MA. Their bias modelling approach allows decisions to be based on all available evidence with less rigorous or less relevant studies getting smaller weights. Their expectation is that bias adjustment will remove much of the heterogeneity in an MA. Bias adjustment is based, however, on elicited opinions rather than empirical evidence.

Differences in study quality may lead to heterogeneity in findings across studies. Ahn and Becker [RR17] compared inverse-variance weighting with weights composed from quality scores on the estimated mean effect in an MA. They conclude that quality weighting adds bias in many cases. They prefer to model the effects of components of quality rather than use quality-score weights.

Yuan and Little [RR18] note that the DerSimonian-Laird (DL) estimate for heterogeneity in an RE MA is in general biased when the patient attrition rate depends on the study-specific effect size. Higher completion rates are associated with more extreme effect sizes, i.e. more bias. They propose three methods to correct for this bias, two of which, the reweighted Bayesian RE model and the Bayesian shared-parameter model work well.

Statistical heterogeneity and small-study effects may affect the validity of an MA. Small-study effects can arise from publication bias and result in a trend to show larger treatment effects for smaller studies in an MA [OR11]. Small-study effects can be seen as a particular case of heterogeneity. To adjust treatment effect estimates for this heterogeneity, Rücker et al [RR19] introduce the limit MA as a new RE model-based method which leads to shrunken, empirical Bayes estimators. This gives rise to a new measure of heterogeneity, G^2, i.e. the proportion of heterogeneity unexplained after allowance for possible small-study effects in the limit MA.

Rare events

An MA on dichotomous outcome data traditionally pools the summary measures of the individual RCTs (e.g. log(odds ratio) or log(risk ratio)) and their standard errors. This assumes an approximately Normal within-study likelihood with known standard errors, does not account for correlation between the estimate and its standard error and necessitates the use of an (arbitrary) continuity correction in case of zero events. To overcome these drawbacks, Stijnen et al [RR20] propose an exact likelihood approach within a generalized linear mixed model. This approach is especially advantageous for sparse (event) data.

Lane [RR21] also notes the limitations of the traditional pooling methods and especially for trials with rare events that are in general not primary outcomes, such as safety outcomes. For these trials, results from an MA should be regarded as only exploratory and hypothesis-generating, in particular when there is much heterogeneity between the trials.

Naïve pooling of cumulative proportions of adverse effects can suffer from Simpson’s paradox when randomization ratios are not identical across studies. Chuang-Stein and Beltangady [RR22] discuss three approaches to report these cumulative proportions of safety data. The inverse sample variance weighting is not recommended; Cochran-Mantel-Haenszel weighting and a study size based method produce similar results.

Gruber and Van der Laan [RR23] compared several estimators of the treatment effect on safety outcomes in an MA for various missingness mechanisms. Their targeted maximum likelihood estimator is asymptotically efficient and unbiased and has good finite sample performance, also when outcomes are missing at random or missingness is informative.
Bennett et al [RR24] compared the standard Cox proportional hazards (PH) model to the Firth penalized Cox PH model and to a Bayesian PH model in MAs with survival-type rare event outcome data. They conclude that the Firth model gives less biased estimates of the (log) hazard ratios than the other two models in rare events survival data.

Series of trials

Chambers et al [RR25] investigated the inclusion of both RCTs and case series in an SR of a rapidly developing technology. Results from non-randomized controlled clinical trials were also included as case series. The authors found no systematic differences in the primary outcome between RCTs and case series and concluded that the evidence base of an SR can be increased and its credibility strengthened by the inclusion of case series. However, they note some clear drawbacks, such as the absence of a control group and several forms of possible bias.
Hee and Stallard [RR26**]** propose a hybrid approach to optimally design an entire development plan encompassing phase II and phase III trials by combining Bayesian decision-theoretic elements and frequentist methods. The phase II trials are assumed to be conducted fully sequentially, i.e. interim decision-making after observation of each new patient, and based on a Bayesian cost-utility approach. From the phase II trials, the most promising treatment is identified and evaluated further in a phase III setting. At the design stage, a prior distribution is assumed for the parameters corresponding to the treatment effects for the experimental treatment. The proposed method assumes that the phase II and III trials have the same patient population, primary endpoint and treatment period.
In the context of a rare disease, often the sample size is retrofitted by adapting the desired power and the relevant effect size to the available number of participants. Le Deley et al [RR27] extended the work of Sposto and Stram [OR12**]** to evaluate the efficiency of a series of successive phase III SCTs by performing an extensive simulation study. Parameters for the simulations were, amongst others, the significance level α, the number and size of trials and the effect size; each trial's outcome was of survival type. When the number of available patients is small, results indicate that designs using smaller sample sizes together with relaxed α values yield greater expected survival benefits. The authors assumed that treatment aspects are similar over trials, that many drugs are available for testing and they did not consider interim analyses.

Multivariate outcomes

A multivariate MA of multiple correlated endpoints enables to borrow strength across the endpoints and to calculate joint confidence and prediction intervals [RR28]. When only aggregate data (AD) of studies to be pooled are available, an estimate for the correlation between the endpoints within a study is necessary. Riley [RR28] shows that ignoring this within-study correlation leads to inaccurate pooled estimates in a bivariate RE MA. Only when between-study variation is very large relative to within-study variation, within-study correlation can be ignored. In general, availability of IPD for all studies to be pooled is desirable. When both IPD and AD are available, a distribution for the correlation can be estimated from the IPD and used as an informative prior distribution for the missing correlations from the AD. Otherwise, sensitivity analyses over a range of values for within-study correlations can be performed. As an alternative, a model with an overall correlation estimate has been proposed by the same author [OR13].

Jackson et al [RR29], commentaries [RR30-34] and the rejoinder [RR35] provide a summary of a one day event on 'Multivariate meta-analysis' for the pooling of studies with multiple, often correlated, outcomes of interest. They discuss the multivariate RE model and its assumptions, describe and apply the estimation methods and discuss advantages and limitations of the multivariate MA. The greatest practical difficulty is the estimation of the within- and between-study correlations, for which the authors describe some solutions. The multivariate Normality assumption is often hard to verify as is the linear relationship of the effects between the studies. Multivariate MA can be useful, but also brings complications and issues. One of the commentaries was that a Bayesian approach using prior information in case of few studies with sparse data can be helpful, but will also show the (large) influence of the prior distribution.

Camilli et al [RR4**]** compared three multi-level meta-regression models for multiple effect sizes per included study, i.e., a standard multi-level model and an iteratively weighted multi-level model, both with weights based on a Normal approximation to the non-central *t* distribution, and a multi-level model based on the exact non-central *t* distribution. The latter model seems to perform better for larger samples. For small samples, however, it is unclear which estimator, a restricted maximum likelihood or a Markov chain Monte Carlo estimator for the between-study variance is better.

Cumulative meta-analysis (CMA)

A CMA evaluates the accumulating evidence of a series of independent RCTs on the same intervention. Mostly, RCTs are included in chronological order into a CMA. Its value, amongst others, lies in the early identification of clinical efficacy or harm, thereby discouraging unnecessary future research. However, periodic updating of MAs can inflate the type I error rate substantially and should be accounted for by formal monitoring procedures [RR36-40,RR10]. Borm et al [RR36] present a rule of thumb that relates the desired type I error and the *P* value of the MA to the maximum number of updates. This rule of thumb does not strictly control the type I error, however.

Trends in effect sizes over time can be detected by visual inspection of cumulative plots or by a test of equality of the estimate of the first RCT and the estimate based on the subsequent RCTs or the overall MA. Bagos and Nikolopoulos [RR41] propose a generalized least squares regression approach to estimate a time trend in effect sizes with a first-order autocorrelation coefficient to adjust for dependence between successive effect size estimates. They applied this exploratory tool in genetic association studies, but also see its usefulness for planning an update of an already published MA.

Sutton et al [RR15] compare two methods to inform prioritization strategies for updating systematic reviews. These methods are only in agreement in case of homogeneity. Although the authors recognize the need to adjust for multiple updating of a CMA, they do not control for this.

Herbison et al [RR42] carried out a number of CMAs to determine the number of trials needed to stable down and get a consistent point estimate. Values for τ^2 and I^2 were no predictors for the number of trials needed nor was the size of the trials. A median of 4 studies were enough to get within 10% of the final point estimate.

Pereira and Ioannidis [RR43] investigated the occurrence of the "*winner's curse phenomenon"*, i.e. the fact that crossing a significance threshold and at the same time estimating the effect size can result in exaggerated effect size estimates, especially for smaller sample sizes. They evaluated a large number of MAs and found that the magnitude of significant effects is often inflated, but the opposite is also true: if a boundary is not crossed, the estimate may be too small. They argue, following other publications, that CMAs should be adjusted for multiple testing.

Trial sequential analysis (TSA)

Like a CMA, a TSA also evaluates the accumulating trial data, but it adjusts for the cumulative updating with O'Brien-Fleming monitoring boundaries. The a priori calculation of the necessary information size for an MA can be performed in a TSA in various ways, amongst others by adjusting for heterogeneity using I^2 [RR37,RR44]. The various information sizes lead to as many sets of trial sequential monitoring boundaries. Thorlund et al [RR44] show that the risk of false-positive results and inaccurate effect size estimates can be reduced by the use of TSA. Brok et al [RR37] find that many published, conclusive MAs are potentially inconclusive when adjusted for the cumulative testing and for heterogeneity. TSA does not allow stopping for futility, however. In a commentary on the previous two papers, Nüesch and Jüni [RR45] emphasize the need for diagnostic measures (such as funnel plots, stratified analyses and interaction tests) to draw conclusions from an MA. Miladinovic et al [RR39] recommend to perform and report sensitivity analyses based on acceptable thresholds for the type I error, power and clinically meaningful treatment difference to prevent premature declaration of a significant MA. They note that three MAs prematurely were declared statistically significant, but later turned out to be not. Imberger et al [RR46] points out that power for two of these three was clearly insufficient to draw a conclusion. Miladinovic et al [RR40] were the first to apply time-to-event TSA. Like the originally proposed TSAs, they did not control for type II error, which made stopping for futility impossible. As an additional comment they note that application of Bayesian monitoring boundaries may result in narrower credibility intervals. For TSAs with count or time-to-event data, software in R and in STATA is presented and described [RR47].

Sequential meta-analysis (SMA)

To guarantee both the type I error and the power of a CMA, an SMA can be implemented using, for example, a triangular test following Whitehead's boundaries approach [OR14]. Van der Tweel and Bollen [RR38] compared TSA and SMA by re-analysing a number of published examples incorporating the Paule-Mandel estimator for heterogeneity between trials in the SMA. They showed that for an SMA (1) no prior estimate for total information size is necessary and thus one set of monitoring boundaries suffices; (2) stopping a CMA for futility is an option; (3) the desired power can be specified in the design; (4) point and interval estimates are adjusted for the multiple testing. The estimates for heterogeneity are, however, unstable for a small number of trials. The paper raised some discussion about supposed differences between TSA and SMA [RR48,RR49].

Novianti et al [RR50**]** evaluated the properties of estimators of heterogeneity in an SMA. Their simulation studies showed that the well-known DL estimator largely underestimates the true value for dichotomous outcomes. They recommend the two-step DL estimator and the Paule–Mandel estimator for use in an SMA with dichotomous or continuous outcomes.

Prospective meta-analysis (PMA)

A PMA can be designed and executed to combine evidence from new and on-going, similar clinical trials in a prospective way. Its advantages are uniformity of the trial protocol, the intervention, the data collection instruments and the reporting of specific outcomes while allowing individual sites some independence with respect to the conduct of research. The inclusion of several sites increases statistical power to address important clinical questions. In PMA, analysis of pooled results is more

facile because of homogeneity of study outcome measures. Besides, IPD enable to conduct stratified analyses and to control for potentially confounding variables. The diversity in study population improves the external validity [RR51]. A PMA is, however, not able to control the generation of new evidence, so the amount, timing and heterogeneity of future trials will not be known in advance. This makes traditional group sequential methods not applicable, but SMA can be applied. [RR52] propose an informative prior distribution to produce a realistic estimate of the between-trial variance in an early stage of an SMA when only a small number of studies is available. The point estimate is then updated in subsequent stages of the SMA. This semi-Bayes approach incorporates the DL estimator. The false-positive and coverage properties depend on the choice of prior distribution for the between-trial variance. Imberger et al [RR53] wonder how the parameters for the prior distribution can be interpreted and how heterogeneity is incorporated.
Shuster and Neu [RR54] argue that prospective group sequential MA methods (such as TSA and SMA) need four essential qualities, i.e. the population effect sizes should be allowed to change over time, independent increments of information from analysis to analysis, robustness against incorrect specification of the information fraction and a physically interpretable effect size. To meet these needs, they impose a separate prior distribution on the effect sizes for each trial, weigh each trial only by sample size and not by the inverse of the variance and apply Pocock's approach to group sequential testing (i.e. a constant nominal type I error probability at each interim analysis). There is no guarantee of power of the PMA, however.
For a recent, practical application of an IPD PMA see Askie et al [OR15].

Network meta-analysis (NMA)

A single SR or MA of a treatment comparison for a single outcome offers a limited view if there are many treatments or many important outcomes to consider. An umbrella review assembles together several SRs on the same condition. If treatments in the SRs can be connected directly or indirectly in a network, outcomes can be analysed with a multiple treatment MA or MTC MA or network MA (NMA). These analyses can also rank the effectiveness of the treatments in a network, thereby determining the best available treatment. An important issue in an NMA is to examine whether there is incoherence or inconsistency, i.e. whether the effect estimated from an indirect treatment comparison differs from that estimated from direct comparisons. However, the power to detect incoherence is low when there are only a few SCTs. Ioannidis [RR55] provides key features in the critical reading of umbrella reviews and key considerations for MT MA. MT MA requires more sophisticated statistical expertise than simple umbrella reviews, but assumes that all data can be analysed together. Most methods for MT MA follow a Bayesian approach. We refer to our nascent review paper on Bayesian methods for a more elaborated discussion on NMA.
Some frequentist approaches to NMA are described [RR56-60]. Piepho et al [RR56] compared the classical (unconditional) two-way model with fixed main effects for trial and treatment with a baseline contrast (conditional) model. They showed the two-way model is simpler to fit than the baseline contrast model and the analysis with the baseline contrast model is not generally invariant to change from baseline. They state that heterogeneity can be separated from inconsistency when there are several trials per trial type, because heterogeneity is a property of variation among trials within the same trial type, whereas inconsistency affects variation between trial types. Stijnen et al [RR20] applied their exact likelihood approach (see also above under General MA) in an example on NMA. Thorlund and Mills [RR57] propose flexible methods for estimating the sample size or

statistical information and the power in an NMA with both direct and indirect treatment comparisons. Their sample size formulas correct for heterogeneity using I^2.

To assess the effect of a particular combination of drugs, Thorlund and Mills [RR58] propose an MTC MA model with an additive-effect parameter. Such a model gains precision by assuming full additivity of treatment effects, that is: when the effect of the treatment combination is equal to the sum of the stand-alone effects. The additive-effects model is superior to the conventional MTC MA model when full additivity holds. The two models are comparably advantageous (in terms of a bias-precision trade-off) when additivity is mildly violated. When additivity is strongly violated, the additive effects model is statistically inferior. When additivity can be assumed, it seems reasonable to prefer the additive effects MTC MA model above the conventional model.

An NMA assumes similarity across the pooled set of trials in terms of patient population and trial characteristics. Naci and O'Connor [RR59] describe the possible benefits of a prospective NMA, such as access to IPD by regulatory agency statisticians, to evaluate comparative efficacy and safety of more than two drugs. Information from both direct and indirect comparisons from a network of trials can provide (far) more information, especially on safety, than just pairwise MA. They urge researchers, manufacturers and regulators to collaborate on future trial designs and analyses. Regulators having access to IPD could also help to inform patients more completely about new treatments. They note, however, that FDA and EMA might not be allowed to use proprietary information from the marketing application of one drug in the evaluation of another.

Bafeta et al [RR60] performed a methodological review of reports of NMAs. They conclude that essential methodological components of the review process, like conducting a literature search and assessing risk of bias of individual studies, are frequently lacking in the reports. They call for guidelines to improve the quality of reporting and conduct of NMAs.

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

Traditionally, an MA combines evidence from related RCTs based on aggregate study-level data. Increasingly, IPD are used. Piepho et al [RR56] compared the analyses of two different models (see also under NMA) with AD and with IPD. For the analysis with AD, both models yield the same results, whereas with IPD analyses are in general not equivalent. Riley et al [RR61] go into the rationale behind IPD MAs. IPD are not needed if the required AD can be obtained in full from publications. However, IPD MAs are potentially more reliable than AD MAs. Use of IPD can increase the power to detect a differential treatment effect, allows adjustment for covariates on patient-level instead of study-level and is particularly advantageous for time-to-event data. A disadvantage is that the IPD approach can take lots of time and costs, and often requires advanced statistical expertise (like FE and RE MA) to preserve the clustering of patients within studies. Increasing use of PMA on IPD is advocated.

To identify a possible source of treatment effect heterogeneity, a treatment-covariate interaction (with the covariate defining the subgroups of interest) can be estimated from a regression analysis on IPD. Kovalchik [RR62] presents an AD EM-algorithm that is equivalent to the maximum likelihood estimates for an IPD linear RE MA with a patient-level treatment-covariate interaction term for a categorical covariate, when the model's variance parameters are known. The presented methodology does not replace an IPD MA, but provides a good AD approximation to a specific kind of IPD interaction model when patient-level data cannot be obtained.

Discussion

Research in rare diseases faces two problems. First, a small number of participants available per trial, and second, usually only a small number of trials targeting the same (new) treatment is possible. In this review we described statistical methods to combine results of series of trials, as published in a recent period of five years. Various search engines were explored. This is specifically important for a methodological review. For example, with the extension to Scopus, 39 unique papers were identified. In total, 62 papers were included in this review. We categorized the relevant methodology according to the type of (meta-)analysis and assessed its usefulness and limitations in small populations.

The focus of this review is on methodology. Completeness is less of an issue in methodological research. A more extensive search could identify additional papers, but is unlikely to provide new insights. In other words, our search will reach a stage (*'theoretical saturation'*) where identifying more articles will not render further methodological perspectives [OR16].

In general, an MA is a well-accepted way of pooling results from a series of trials. Various approaches to MA have been described and evaluated in the past. Herbison et al [RR42] concluded that a median number of 4 studies are needed to get within 10% of the final pooled point estimate, where they based this final value on a minimum of 10 trials and assumed it the true value. They restricted themselves to FE estimates based on 95% CIs and did not adjust the CIs for multiple testing. They recognize that it is impossible to predict which SRs with a small number of studies will be correct in the long run.

Simulation studies with survival outcomes showed that designs using smaller sample sizes and relaxed α values yield greater expected survival benefits than traditional design strategies that aimed to detect a small difference with high level of evidence [RR27] with reference to Sposto and Stram [OR12]. These studies focused on personalized medicine, but can also be useful for RCTs in rare diseases. Research has to confirm the results for dichotomous and continuous outcomes. O'Connor and Hemmings [OR3] also suggested relaxation of the type I error.

Both Miladinovic et al [RR39] and Nüesch and Jüni [RR45] cite Egger and Davey Smith [OR17] that 'results of meta-analyses that are exclusively based on small trials should be distrusted - even if the combined effect is statistically highly significant. Several medium-sized trials of high quality seem necessary to render results trustworthy.' This citation is opposite to the suggestion by IntHout et al [OR18] that 'evidence of efficacy based on a series of smaller trials may lower the error rates compared with a single well-powered trial'.

Most research in MA acknowledges the need to incorporate heterogeneity into the effect point- and interval estimates. The properties of the estimators are not well-known though for a small number of trials. Various authors note that both I^2 and τ^2 as measures for heterogeneity can be unreliable and unstable in an MA with a small number of trials. Estimating heterogeneity is considered more important than testing it. To account for the uncertainty in the estimated value of τ^2 in a CI for the pooled effect size, the use of a *t*-distribution with k-1 or k-2 degrees of freedom (with k the number of trials pooled) instead of a Normal distribution in a CI for the pooled effect size is proposed [RR5,RR8,RR39,RR45].

For continuous outcomes a variance stabilizing transformation is advised [RR8] before estimating the confidence interval. The DL method-of-moments estimator to estimate the between-study heterogeneity parameter τ^2 is widely applied [OR7,RR7,RR8,RR16] and is also standard in most software. Yuan and Little [RR18] observe a bias in the DL estimator leading to too narrow CIs. Turner et al [RR16] and Novianti et al [RR50] note that alternative estimators such as proposed by

DerSimonian and Kacker [OR19] might be preferred. Their properties, and those of other recently proposed estimators [OR20], in a small number of SCTs have to be explored.

Case series of the use of therapeutic procedures or devices can be included to strengthen the evidence in an SR, although Chambers et al [RR9] mention some drawbacks. The contribution of case series of drug use for an SR and MA in rare diseases has to be further explored.

Hee and Stallard [RR26] propose an optimal decision-theoretic design of a series of phase II clinical trials followed by a phase III RCT. Their approach is a hybrid one, in that it assumes prior distributions for the success probabilities in the phase II trials, followed by a classical frequentist hypothesis test. This proposal can be useful in rare diseases, but its application in RCTs with non-dichotomous outcomes has to be investigated further.

Frequently, an MA is updated with results of one or more newly published RCTs, leading to a so-called CMA. In general, such an update does not control for multiple testing, thereby risking an increase in the overall type I error. A TSA or SMA design, on the contrary, guarantees the overall type I error. The use of a TSA, an SMA or a PMA enables to stop a series of trials for efficacy or futility, thereby leading to efficiency gains and thus ethical and/or economic benefits. Ideally, TSA should be applied prospectively with clinically relevant pre-specified treatment differences, type I and type II errors [RR39]. These authors also see a role for sensitivity analyses.

Thorlund and Mills [RR56] use I^2 to correct for heterogeneity in an NMA. Its use as a measure of heterogeneity is, however, debated. It is, for example, known to increase with the number of patients included in the studies in a MA [OR21]. Wetterslev et al [RR14] conclude that their proposed measure D^2 seems a better alternative for trial diversity and for adjustment of the required information size. Moreover, it adapts automatically to different between-trial variance estimators, while I^2 is linked to the DL estimator. Demidenko et al [RR7] developed a coefficient of determination to measure the strength of the presence of random effects in a model. It is unclear what its additional value is to I^2. Von Hippel [OR22] showed that I^2 is imprecise and biased in small meta-analyses and emphasizes its cautious interpretation and presentation in small MAs.

Higgins et al [RR52] consider clinical research a sequential process where SMA can play a role in the design of a new trial, since the amount of further information that would be required can be determined. They notice, however, also some points of attention. One is whether or not a correction for multiple looks to cumulative data is needed. Another is the poor estimation of τ^2 from a small number of studies. Then realistic prior information is necessary, but the choice of the prior distribution is crucial in the early stages of an SMA. Undertaking an MA in a fully Bayesian way has the advantage that no correction for multiple looks is necessary for inference, but frequentist properties, such as type I errors, can be inflated.

Rücker et al [RR19] suggest to adjust treatment effect estimates for small-study effects, leading to shrunken, empirical Bayes estimates. These estimates are approximately unbiased when the number of trials in an MA is at least 10. The approach depends on the estimator for τ^2, which was the DL estimator, which is known to underestimate τ^2 for dichotomous outcomes. The remaining amount of heterogeneity, termed G^2, varies considerably depending on the estimator used. This approach should only be used in an MA with more than 10 trials with one or more medium or large-sized trials and clear variation in trial size [RR19,OR11].

Especially in rare diseases, multiple outcomes will (have to) be examined simultaneously. In that case a multivariate MA as proposed by Jackson et al [RR29-35] may show potential, but also raises concerns. In particular, the statistical properties for a small number of small samples, imprecise between-study (co)variances, unavailable within-study correlation estimates, a possible large number of parameters to be estimated, and missing outcomes in some but not all trials require further study. It also makes clear that IPD will have to be available for RCTs in such MAs. Riley [RR28] also points to the important role of a multivariate MA in evidence-based decision making. His approach assumes the within-study correlation as given and known, though. Comparison of this approach with an earlier proposed alternative [OR13], a model with an overall correlation estimate, in small populations deserves further investigation. Stijnen et al [RR20] presented an extension of their exact likelihood method for dichotomous outcomes into a multivariate MA. Their model can also be applied with rare event outcomes.

Both frequentist and Bayesian approaches are applied to combine successfully the extracted data from several trials. Their application in the field of rare diseases is one possible way to sufficiently support a treatment effect. Measures for heterogeneity can be unreliable and unstable in an MA based on a small number of trials. An option is to formulate an informative prior distribution around τ^2. This prior can be updated in an SMA with the result of a new MA leading to a posterior distribution, which in turn forms a new prior. However, for small data, Bayesian posterior probabilities may depend heavily on the choice of the prior distribution. Higgins et al [RR52] prefer a Bayesian approach, especially for prediction. They note, however, that their approach does not lend itself well to rare events. Furthermore, it is not clear that strict control over false-positive findings is important in this context, since a small, non-statistically significant, signal should still be investigated when the adverse effect is major. Both Higgins et al [RR5] and Chung et al [RR6] prefer a Bayesian informative prior distribution for the heterogeneity parameter. The proposed Bayes modal estimator prevents zero (i.e. boundary) estimates and shows good properties for a small number of studies. Most methods for NMA follow a Bayesian approach. Ioannidis [RR55] notes that the power to detect incoherence in an NMA is low when the network consists of only a few (small) trials.

In general, availability of IPD for all studies to be pooled is desirable, particularly in rare diseases. There, the role of regulators is a further point of attention, because of the remark made by Naci and O'Connor [RR58] that FDA and EMA might not be allowed to use proprietary information from the marketing application of one drug in the evaluation of another.

Recently, several initiatives have been started to facilitate and promote the sharing of clinical trial data. Members of three EU-FP7 projects on small populations (Asterix, Ideal and Inspire) together with representatives from regulatory agencies, scientific journals and industry addressed the arising intricate biostatistical questions such as the interpretation of multiple statistical analyses, both prospective and retrospective as well as the issue of data protection which is most prominent in the setting of rare diseases [OR23].

Conclusions

Our review covered a five-year period. We noticed that in this period only few papers pay attention to a small series of small trials. This finding makes further research necessary. For evidence-based decision-making on a (small) number of trials in small populations, we see several directions for further investigation: 1) frequentist properties of estimators for heterogeneity between trials; 2) use of exact (likelihood) methods; 3) value of prospective meta-analysis in drug development; 4) combination of observational, historical and trial data to ensure that every patient contributes as much information as possible [OR3,OR24]; 5) relax the type I error probability; 6) focus on multiple outcomes per patient; 7) combination of IPD with AD; 8) special attention for the evaluation of rare events, such as safety outcomes.

Abbreviations

AD: aggregate study-level data
CI: confidence interval
CMA: cumulative meta-analysis
DL: DerSimonian-Laird
FE: fixed effect
IPD: individual patient data
MA: meta-analysis
MT: multiple treatments
MTC: mixed treatment comparison
NMA: network meta-analysis
PH: proportional hazards
PMA: prospective meta-analysis
RCT: randomized controlled trial
RE: random effects
SCT: small clinical trial
SMA: sequential meta-analysis
SR: systematic review
TSA: trial sequential analysis

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors together conceived the review, IvdT and KP devised the search terms, KP performed the literature search, KP, IvdT and CvB discussed the in- and exclusion of papers and IvdT drafted the manuscript. All authors provided feedback on the draft versions and read and approved the final manuscript.

Acknowledgements

We thank Stavros Nikolakopoulos for helpful comments on the manuscript.

This work was supported by the EU FP7 HEALTH.2013.4.2-3 project "Advances in Small Trials dEsign for Regulatory Innovation and eXcellence" (Asterix): Grant nr 603160.

Appendix: Search strategy

search engine	search query	nr of results
PubMed	((((trial[text] OR trials[text] OR "clinical trials as topic"[MeSH Major topic]) AND (meta*analysis[Title/Abstract] OR "meta analysis as topic"[MeSH Major topic]) AND ("Mixed treatment"[Title/Abstract] OR Indirect[Title] OR prospective [Title/Abstract] OR network [Title/Abstract] OR cumulative [Title/Abstract] OR sequential [Title/Abstract] OR bayes* [Title/Abstract] OR "hierarchical mode*" [Title/Abstract])) OR (series[Title] AND trials[Title] AND "clinical trials as topic"[MeSH Major Topic]))) NOT ("phase I"[title/abstract] OR "phase IV"[title/abstract] OR "a systematic review"[title] OR "a meta-analysis"[title])) AND ("2009/1/1"[Date - Publication] : "2013/12/31"[Date - Publication])	**1031**
Scopus	(ALL(trial OR trials)AND TITLE-ABS-KEY(meta*analysis OR meta-analysis OR "meta analysis")AND TITLE-ABS-KEY(prospective OR network OR cumulative OR sequential OR bayes* OR bayes OR "hierarchical models") AND NOT TITLE-ABS-KEY("phase I" OR "phase IV" OR "a systematic review"OR "a meta-analysis") OR TITLE-ABS-KEY(series AND trials AND "clinical trials")) AND PUBYEAR > 2008 PUBYEAR < 2014	**2438**
Web of Science	TOPIC: (trial OR trials) AND TOPIC : ("meta-analysis" OR "meta analysis") AND TOPIC: (prospective OR network OR cumulative OR sequential OR bayes* OR bayes OR "hierarchical models") NOT TITLE: ("phase I" OR "phase IV" OR "a systematic review" OR "a meta-analysis") OR TOPIC: (series AND trials AND "clinical trials") Timespan=2009-2014. Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH.	**2230**

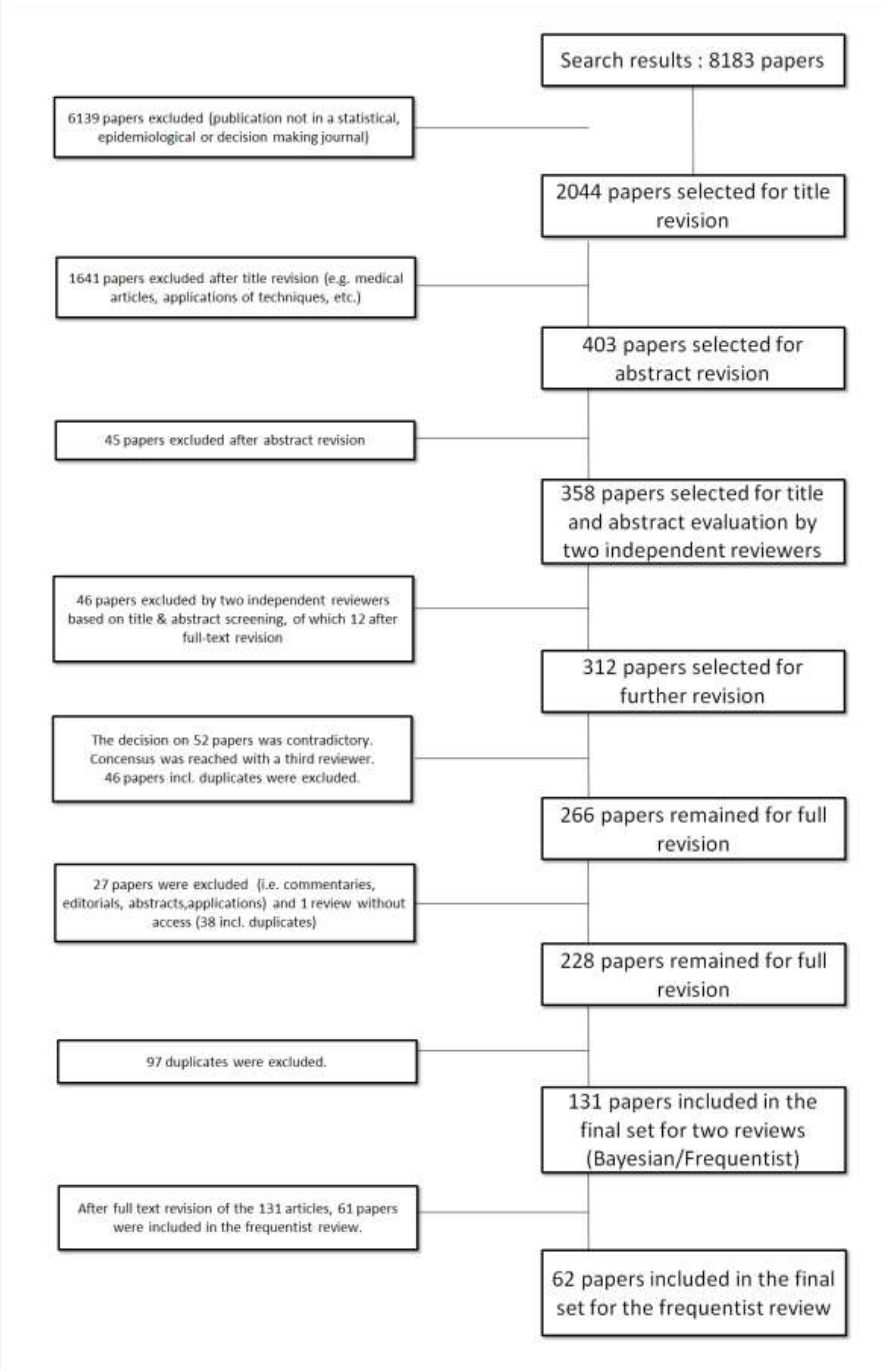
JSTOR	(((((trial OR trials) AND ("meta-analysis" OR "meta analysis")) AND (prospective OR network OR cumulative OR sequential OR bayes* OR "hierarchical models")) NOT ("phase I" OR "phase IV" OR "a systematic review" OR "a meta-analysis")) OR (series AND trials AND "clinical trials")) AND ((year:2009 AND month:01 AND day:[01 TO 31]) OR (year:2009 AND month:[02 TO 12]) OR (year:[2010 TO 3000]))	**2436**
Cochrane	trial or trials and "meta-analysis" or "meta analysis":ti,ab,kw and prospective or network or cumulative or sequential or bayes* or bayes or "hierarchical models":ti,ab,kw or "a systematic review" or series and trials and "clinical trials":ti not "phase I" or "phase IV" or "a systematic review" or "a meta-analysis":ti from 2009 to 2013	**48**/556, considering only methodological papers

Table 1. Characteristics of the reviewed papers

Characteristic	Category*	n
Year of publication	2009	15
	2010	10
	2011	15
	2012	9
	2013	13
Type of MA	General MA	37
	CMA	14
	TSA	13
	SMA	6
	PMA	9
	NMA / MTC	9
	case series	3
Input data	AD	49
	IPD	7
	AD & IPD	6
Outcome	univariate	51
	multivariate	8
	uni- and multivariate	3
Type of outcome	dichotomous	42
	continuous	14
	time-to-event	16
	other	2
	n.s.	10
Approach	frequentist	47
	frequentist and Bayes	9
	hybrid	5
	empirical Bayes	1
Total nr of papers		62

*MA = meta-analysis, CMA = cumulative meta-analysis, TSA = trial sequential meta-analysis, SMA = sequential meta-analysis, PMA = prospective meta-analysis, NMA = network meta-analysis, MTC = mixed treatment comparison, AD = aggregated data, IPD = individual patient data, n.s.= not specified.

Figure 1. Flow diagram of the search strategy


Search results : 8183 papers
6139 papers excluded (publication not in a statistical, epidemiological or decision making journal)
2044 papers selected for title revision
1641 papers excluded after title revision (e.g. medical articles, applications of techniques, etc.)
403 papers selected for abstract revision
45 papers excluded after abstract revision
358 papers selected for title and abstract evaluation by two independent reviewers
46 papers excluded by two independent reviewers based on title & abstract screening, of which 12 after full-text revision
312 papers selected for further revision
The decision on 52 papers was contradictory. Concensus was reached with a third reviewer. 46 papers incl. duplicates were excluded.
266 papers remained for full revision
27 papers were excluded (i.e. commentaries, editorials, abstracts,applications) and 1 review without access (38 incl. duplicates)
228 papers remained for full revision
97 duplicates were excluded.
131 papers included in the final set for two reviews (Bayesian/Frequentist)
After full text revision of the 131 articles, 61 papers were included in the frequentist review.
62 papers included in the final set for the frequentist review

References of papers in the review (RR)

1. Aiello F, Attanasio M, Tinè F. Assessing covariate imbalance in meta-analysis studies. Stat Med 2011; 30: 2671–2682.
2. Verbeek J, Ruotsalainen J, Hoving JL. Synthesizing study results in a systematic review. Scand J Work Environ Health 2012; 38: 282–290.
3. Greenland S. Accounting for uncertainty about investigator bias: disclosure is informative. J Epidemiol Community Health 2009; 63: 593–598.
4. Camilli G, de la Torre J, Chiu CY. A noncentral t regression model for meta-analysis. J Educ Behav Stat 2010; 35: 125–153.
5. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc 2009; 172: 137–159.
6. Chung Y, Rabe-Hesketh S, Choi I-H. Avoiding zero between-study variance estimates in random-effects meta-analysis. Stat Med 2013; 32: 4071–4089.
7. Demidenko E, Sargent J, Onega T. Random effects coefficient of determination for mixed and meta-analysis models. Commun Stat Theory Methods 2012; 41: 953–969.
8. Malloy MJ, Prendergast LA, Staudte RG. Transforming the Model T: random effects meta-analysis with table weights. Stat Med 2013; 32: 1842–1864.
9. Viechtbauer W. Conducting Meta-Analyses in R with the metafor package. J Stat Softw 2010; 36: 48.
10. Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. BMC Med Res Methodol 2009; 9: 29.
11. Goudie AC, Sutton AJ, Jones DR, Donald A. Empirical assessment suggests that existing evidence could be used more fully in designing randomized controlled trials. J Clin Epidemiol 2010; 63: 983–991.
12. Ioannidis J, Karassa F. The need to consider the wider agenda in systematic reviews and meta-analyses. BMJ 2010; 341: 762–765.
13. Rotondi M, Donner A. Sample size estimation in cluster randomized trials: An evidence-based perspective. Comput Stat Data Anal 2012; 56: 1174–1187.
14. Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC Med Res Methodol 2009; 9: 86.
15. Sutton AJ, Donegan S, Takwoingi Y, Garner P, Gamble C, Donald A. An encouraging assessment of methods to inform priorities for updating systematic reviews. J Clin Epidemiol 2009; 62: 241–251.
16. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. J R Stat Soc A 2009; 172: 21–47.
17. Ahn S, Becker BJ. Incorporating quality scores in meta-analysis. J Educ Behav Stat 2011; 36:555–585.
18. Yuan Y, Little RJA. Meta-analysis of studies with missing data. Biometrics 2009; 65: 487–496.
19. Rücker G, Schwarzer G, Carpenter JR, Binder H, Schumacher M. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. Biostatistics 2011; 12: 122–142.
20. Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Stat Med 2010; 29: 3046–3067.
21. Lane PW. Meta-analysis of incidence of rare events. Stat Methods Med Res 2013; 22: 117–132.
22. Chuang-Stein C, Beltangady M. Reporting cumulative proportion of subjects with an adverse event based on data from multiple studies. Pharm Stat 2011; 10: 3–7.
23. Gruber S, van der Laan MJ. An application of targeted maximum likelihood estimation to the meta-analysis of safety data. Biometrics 2013; 69: 254–262.
24. Bennett MM, Crowe BJ, Price KL, Stamey JD, Seaman JW. Comparison of Bayesian and frequentist meta-analytical approaches for analyzing time to event data. J Biopharm Stat 2013; 23: 129–145.

25. Chambers D, Rodgers M, Woolacott N. Not only randomized controlled trials, but also case series should be considered in systematic reviews of rapidly developing technologies. J Clin Epidemiol 2009; 62: 1253–1260.
26. Hee SW, Stallard N. Designing a series of decision-theoretic phase II trials in a small population. Stat Med 2012; 31: 4337–4351.
27. Le Deley M-C, Ballman KV, Marandet J, Sargent D. Taking the long view: how to design a series of Phase III trials to maximize cumulative therapeutic benefit. Clin Trials 2012; 9: 283–292.
28. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. J R Stat Soc A 2009; 172: 789–811.
29. Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. Stat Med 2011; 30: 2481–2498.
30. Hedges LV. Comment on: 'Multivariate meta-analysis: potential and promise'. Stat Med 2011; 30: 2499.
31. Cox DR. Multivariate meta-analysis: a comment. Stat Med 2011; 30: 2500-2501.
32. Bland JM. Comments on: 'Multivariate meta-analysis: potential and promise'. Stat Med 2011; 30: 2502-2503.
33. Gasparrini A, Armstrong B. Multivariate meta-analysis: a method to summarize non-linear associations. Stat Med 2011; 30: 2504-2506
34. Harbord RM. Commentary on: 'Multivariate meta-analysis: potential and promise'. Stat Med 2011; 30: 2507-2508.
35. Jackson D, White IR, Riley RD. Rejoinder to commentaries on "Multivariate meta-analysis: Potential and promise." Stat Med 2011; 30: 2509–2510.
36. Borm GF, Donders RT. Updating meta-analyses leads to larger type I errors than publication bias. J Clin Epidemiol 2009; 62: 825–830.
37. Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive--Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. Int J Epidemiol 2009; 38: 287–298.
38. Van der Tweel I, Bollen C. Sequential meta-analysis: an efficient decision-making tool. Clin Trials 2010; 7: 136–146.
39. Miladinovic B, Kumar A, Hozo I, Mahony H, Djulbegovic B. Trial sequential analysis may be insufficient to draw firm conclusions regarding statistically significant treatment differences using observed intervention effects: a case study of meta-analyses of multiple myeloma trials. Contemp Clin Trials 2013; 34: 257–261.
40. Miladinovic B, Mhaskar R, Hozo I, Kumar A, Mahony H, Djulbegovic B. Optimal information size in trial sequential analysis of time-to-event outcomes reveals potentially inconclusive results because of the risk of random error. J Clin Epidemiol 2013; 66: 654–659.
41. Bagos PG, Nikolopoulos GK. Generalized least squares for assessing trends in cumulative meta-analysis with applications in genetic epidemiology. J Clin Epidemiol 2009; 62: 1037–1044.
42. Herbison P, Hay-Smith J, Gillespie WJ. Meta-analyses of small numbers of trials often agree with longer-term results. J Clin Epidemiol 2011; 64: 145–153.
43. Pereira T V, Ioannidis JPA. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. J Clin Epidemiol 2011; 64: 1060–1069.
44. Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JPA, Thabane L, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2009; 38: 276–286.
45. Nüesch E, Jüni P. Commentary: Which meta-analyses are conclusive? Int J Epidemiol 2009; 38: 298–303.
46. Imberger G, Wetterslev J, Gluud C. Trial sequential analysis has the potential to improve the reliability of conclusions in meta-analysis. Contemp Clin Trials 2013; 36: 254–255.
47. Miladinovic B, Hozo I, Djulbegovic B. Trial sequential boundaries for cumulative meta-analyses. The Stata J 2013; 13: 77–91.

48. Thorlund K, et al. Comments on “Sequential meta-analysis: an efficient decision-making tool” by I van der Tweel and C Bollen. Clin Trials 2010; 7: 752–753.
49. Van der Tweel I, Bollen C. Response to Letter from K Thorlund, et al. Clin Trials 2010; 7: 754.
50. Novianti PW, Roes KCB, van der Tweel I. Estimation of between-trial variance in sequential meta-analyses: A simulation study. Contemp Clin Trials 2014; 37: 129–138. With a corrigendum in Contemp Clin Trials 2015; 41: 335.
51. Turok DK, Espey E, Edelman AB, Lotke PS, Lathrop EH, Teal SB, et al. The methodology for developing a prospective meta-analysis in the family planning community. Trials 2011; 12: 104.
52. Higgins JPT, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. Stat Med 2011; 30: 903–921.
53. Imberger G, Gluud C, Wetterslev J. Comments on “Sequential methods for random-effects meta-analysis” by JPT Higgins, et al. Stat Med 2011; 30: 2965–2966.
54. Shuster JJ, Neu J. A Pocock approach to sequential meta-analysis of clinical trials. Res Synth Methods 2013; 4: 269-279.
55. Ioannidis JPA. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. CMAJ 2009; 181: 488–493.
56. Piepho HP, Williams ER, Madden LV. The use of two-way linear mixed models in multitreatment meta-analysis. Biometrics 2012; 68: 1269-1277.
57. Thorlund K, Mills EJ. Sample size and power considerations in network meta-analysis. Syst Rev 2012; 1, 41.
58. Thorlund K, Mills E. Stability of additive treatment effects in multiple treatment comparison meta-analysis: a simulation study. Clin Epidemiol 2012; 4: 75-85.
59. Naci H, O’Connor AB. Assessing comparative effectiveness of new drugs before approval using prospective network meta-analyses. J Clin Epidemiol 2013; 66: 812–816.
60. Bafeta A, Trinquart L, Seror R, Ravaud P. Analysis of the systematic reviews process in reports of network meta-analyses: methodological systematic review. BMJ 2013; 347: 1-12.
61. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010; 340: 521-525.
62. Kovalchik SA. Aggregate-data estimation of an individual patient data linear random effects meta-analysis with a patient covariate-treatment interaction term. Biostatistics 2013; 14: 273–283.

Other references (OR)

1. Asterix project: http://www.asterix-fp7.eu/ Accessed 12 Oct 2015.
2. O'Connor DJ, Hemmings RJ. Coping with small populations of patients in clinical trials. Expert opinion on Orphan drugs 2014; 2: 765-768.
3. http://ec.europa.eu/health/ph_information/documents/ev20040705_rd05_en.pdf. Accessed 12 Oct 2015.
4. CHMP/EWP/83561/2005. Guideline on clinical trials in small populations. London: EMEA, 2006. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003615.pdf. Accessed 12 Oct 2015.
5. Gupta S, Faughnan ME, Tomlinson GA, Bayoumi AM. A framework for applying unfamiliar trial designs in studies of rare diseases. J Clin Epidemiol 2011; 64: 1085-1094.
6. Cornu C, Kassai B, Fisch R, Chiron C, Alberti C, Guerrini R, et al and the CRESim & Epi-CRESim Project Groups. Experimental designs for small randomized clinical trials: an algorithm for choice. Orphanet Journal of Rare Diseases 2013; 8: 48
7. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. United Kingdom: John Wiley & Sons; 2009.
8. National Research Council. Combining information: Statistical issues and opportunities for research. Washington, DC: National Academy Press, 1992.
9. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21: 1539-1558.
10. Hardy R, Thompson SG. A likelihood approach to meta-analysis with random effects. Stat Med 1996; 15: 619–629.
11. Sterne JAC, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the statistical literature. J Clin Epidemiol 2000; 53: 1119-11129.
12. Sposto R, Stram DO. A strategic view of randomized trial design in low-incidence paediatric cancer. Stat Med 1999; 18: 1183–1197.
13. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. Biostatistics 2008; 9: 172-186.
14. Whitehead J. The Design and Analysis of Sequential Clinical Trials (rev. 2nd ed). John Wiley & Sons Ltd, Chichester, UK, 1997.
15. Askie LM, Bauer LA, Campbell K, Daniels LA, Hesketh K, Magarey A, et al; EPOCH Collaboration Group. The early prevention of obesity in children (EPOCH) Collaboration-an individual patient data prospective meta-analysis. BMC Public Health 2010; 10: 728.
16. Lilford RJ, Richardson A, Stevens A, Fitzpatrick R, Edwards S, Rock F, Hutton JL. Issues in methodological research: perspectives from researchers and commissioners. Health Technol Assess 2001; 5: 1–57.
17. Egger M, Davey Smith G. Misleading meta-analysis. BMJ 1995; 310: 752–754.
18. IntHout J, Ioannidis JP, Borm GF. Obtaining evidence by a single well-powered trial or several modestly powered trials. Stat Meth Med Res 2012; 0: 1-15.
19. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemp Clin Trials 2007; 28: 105–114.
20. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Meth 2014; 14: 25.
21. Rücker G, Schwarzer G, Carpenter J, Schumacher M. Undue reliance on I^2 in assessing heterogeneity may mislead. BMC Med Res Meth 2008; 8: 79–87.
22. Von Hippel PT. The heterogeneity statistic I^2 can be biased in small meta-analyses. BMC Med Res Meth 2015; 15:35.

23. Koenig F, Slattery J, Groves T, Lang T, Benjamini Y, Day S, et al. Sharing clinical trial data on patient level: Opportunities and challenges. Biom J 2015; 57: 8-26.
24. Verde PE, Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. Res Synth Meth 2015; 6: 45-62.