Clinical Rheumatology (2019) 38:2029–2038
https://doi.org/10.1007/s10067-019-04590-6

REVIEW ARTICLE

Systematic reviews and meta-analysis in rheumatology: a gentle introduction for clinicians

George A. Kelley[1] · Kristi S. Kelley[1]

Received: 25 March 2019 / Revised: 30 April 2019 / Accepted: 2 May 2019 / Published online: 17 May 2019

Abstract
Given the plethora of studies today on the same topic, clinicians in rheumatology as well as others increasingly rely on systematic reviews, with or without meta-analysis, to aid in their evidence-based decision-making. However, given time constraints, staying up-to-date on current methods for conducting systematic reviews and meta-analyses as well as interpreting the results of these reviews for application in clinical practice can be challenging. The purpose of this paper is to try and address this gap. In this paper, a description of the different types of systematic reviews and meta-analyses is provided as well as a description of the major elements, including methodology and interpretation of systematic reviews with meta-analyses. Included is a broad, five-question checklist to aid clinicians in rheumatology for making decisions about the utility of a systematic review. It is the hopes that this paper will aid clinicians in rheumatology as well as other consumers of systematic reviews and meta-analyses with the information necessary for judging the utility of systematic reviews and meta-analyses in their own work.

Keywords Clinicians · Meta-analysis guide · Rheumatology · Systematic reviews

Introduction

Clinical rheumatologists today traditionally rely on an evidence-based approach in their clinical decision-making that includes their own clinical judgment, patient preference, and evidence derived from research studies. Given the plethora of studies today on the same topic, the number of systematic reviews, with or without meta-analysis, has increased dramatically, including the field of rheumatology (Fig. 1). Given this increase and potential uptake, it is important for clinicians in rheumatology to develop a basic understanding of systematic reviews with meta-analysis so as to best judge the utility of such with respect to both clinical decision-making and possibly their own research. The purpose of this paper is to attempt to address this need. The authors first introduce the reader to the different types of systematic reviews, discuss the different types of meta-analyses that are included in systematic reviews, and, finally, describe the major components of a systematic review with meta-analysis, including the interpretation of results. While there is a focus on traditional aggregate data (AD) systematic reviews with meta-analysis, the most common type published today, much of the information is applicable across all types of systematic reviews with meta-analysis. To simplify reading and clinical application, there is a purposeful avoidance of formulas. Selected references where more details on each respective item may be found are provided throughout the document.

See related article, https://doi.org/10.1007/s10067-019-04666-3

✉ George A. Kelley
gkelley@hsc.wvu.edu

Kristi S. Kelley
kskelley@hsc.wvu.edu

[1] Meta-Analytic Research Group, School of Public Health, Department of Biostatistics, Robert C. Byrd Health Sciences Center, West Virginia University, PO Box 9190, Morgantown, WV 26506-9190, USA

Types of systematic reviews

Generally speaking, reviews may take the form of (1) a *scoping review*; (2) *systematic review of previous systematic reviews* (SRPSR), with or without meta-analysis, also known as "umbrella reviews," "overviews of reviews," "reviews of reviews," "summary of systematic reviews," "synthesis of reviews," or "meta-reviews" [1]; and (3) *systematic reviews with or without meta-analysis*. A *scoping*

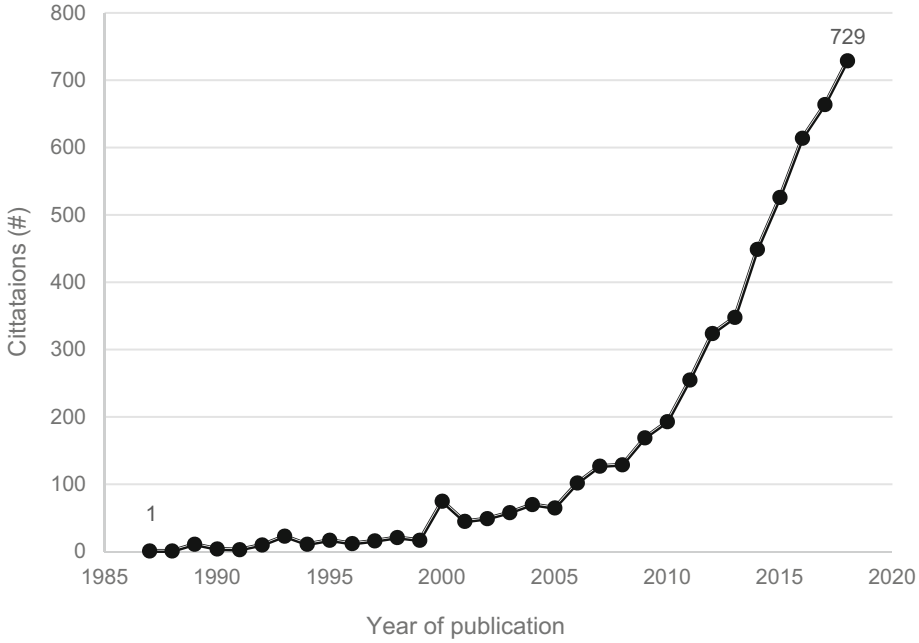

Fig. 1 Trend graph of PubMed search for rheumatology citations. This simple search was conducted in PubMed on February 19, 2018, using the search phrase rheumat* and ("meta-analysis" or "systematic review"). The *x*-axis represents the year while the *y*-axis represents the number of citations for that year. As can be seen, there has been a large increase since 1989, including a steep increase from 2010 to 2018

review may be defined as a type of research synthesis that aims to "map the literature on a particular topic or research area and provide an opportunity to identify key concepts, gaps in the research, and types and sources of evidence to inform practice, policymaking, and research" [2]. For example, Misra et al. recently conducted a scoping review on the use of non-biologic disease-modifying anti-rheumatic drugs (DMARDs) in the management of large vessel vasculitis (LVV) [3]. The authors concluded that (1) methotrexate, leflunomide, azathioprine, mycophenolate mofetil, and cyclophosphamide were effective in Takayasu's arteritis (TA) but the quality of evidence was low; (2) that high-quality evidence for methotrexate and moderate-quality evidence for hydroxychloroquine and cyclosporine suggested that all were ineffective for giant cell arteritis (GCA); and (3) that moderate-quality evidence for azathioprine and low quality of evidence for leflunomide, mycophenolate mofetil, cyclophosphamide, and dapsone suggested that all were effective in GCA [3]. The authors also concluded that a lack of high-quality evidence was currently available to recommend or not recommend the use of conventional disease-modifying anti-rheumatic drugs (cDMARDs) in TA and GCA and that a need exists for large multi-center randomized placebo-controlled trials to examine the use of cDMARDs in LVV [3]. Details for conducting and reporting scoping reviews can be found elsewhere [4].

A number of systematic reviews with or without meta-analysis now exist on the same topic. As a result, there is now a need for *SRPSR* in order to provide a synthesis of the evidence to clinicians and others as well as conducting future research, including information on whether an original or updated systematic review on the topic of interest is needed [5]. Recognition of the importance of this approach is illustrated by a recent thematic series dedicated to SRPSR [6]. As an example of this approach in rheumatology, Christie et al. conducted a SRPSR of 28 systematic reviews on the effectiveness of nonpharmacologic and nonsurgical interventions for patients with rheumatoid arthritis [7]. The overall conclusion was that the quality of evidence regarding the effectiveness of most nonpharmacologic and nonsurgical interventions in rheumatoid arthritis was moderate to low and that additional research was needed [7]. General guidelines for conducting SRPSR may be found elsewhere [8].

The most common types of reviews conducted today are *systematic reviews of individual studies.* These may or may not include a quantitative synthesis, i.e., meta-analysis, of the data. The Cochrane Collaboration defines a systematic review as "a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies that are included in the review. Statistical methods (**meta-analysis**) may or may not be used to analyze and summarize the results of the included studies" [1]. The decision on whether to include a meta-analysis in a systematic review is usually predicated on the author's decision of whether the primary outcome(s) of interest can be combined into one common metric. If not, the data are synthesized qualitatively. However, most systematic reviews today include a quantitative synthesis, i.e., meta-analysis of the data. The remainder of this paper will focus on systematic reviews that include a meta-analysis.

Individual participant data and aggregate data meta-analysis

Generally, there are two types of meta-analyses, an *individual participant data* (IPD) meta-analysis and an *aggregate data* (AD) meta-analysis. Considered the gold standard [9], an IPD meta-analysis pools patient-level data from each individual for analysis. In contrast, an AD meta-analysis pools study-level data from summary statistics on patients, for example, sample sizes, means, and standard deviations, provided in each original study. One of the major advantages of the IPD approach is the ability to analyze data such as age at the level of the individual as opposed to group mean data from an AD meta-analysis, thereby helping to avoid what is known as ecological fallacy, i.e., the possibility that results at the group level do not occur at the patient level [9]. However, major challenges exist with respect to the conduct of an IPD meta-analysis. For example, a previous study of 142 IPD meta-analyses from different subject areas reported that the obtaining of IPD ranged from 25 to 100% [10]. With respect to rheumatology, the authors were only able to retrieve useable IPD from eight of 29 eligible studies (28%) for a meta-analysis on exercise and depressive symptoms in adults with arthritis and other rheumatic diseases [11]. The inability to retrieve IPD from eligible studies is problematic because it can lead to an increased risk of bias. While methods have been proposed for conducting an IPD meta-analysis that allows one to pool both IPD and AD for studies in which IPD cannot be obtained [12], one is still left with AD for one or more studies. Another major challenge of conducting an IPD versus AD meta-analysis is the increased time and resources that are needed. Steinberg et al. estimated that the costs associated with conducting their own IPD meta-analysis were more than 5 times greater using the IPD versus AD approach [13]. However, others estimated the costs for the same study to be at least 8 times greater given that the research team continued to work on the study after funding for the project ended [14]. Detailed guidance on the conduct, use, and reporting of IPD meta-analyses may be found elsewhere [1, 9, 15].

While the number of IPD meta-analyses are increasing, AD meta-analyses are still the most common. For example, a PubMed search of rheumatology articles conducted by the authors on February 22, 2019, and which excluded IPD meta-analyses yielded a total of 5095 citations between 1987 and 2018 (unpublished results). When limited to IPD meta-analyses, the total number of citations was 40 for the same time period (unpublished results). Detailed guidance for conducting and reporting AD meta-analyses have been reported elsewhere [16–18].

Network meta-analyses and non-inferiority meta-analyses

A meta-analysis based on either AD or IPD often compares two treatments at a time, usually an intervention versus control condition from a randomized controlled trial. In contrast, a *network meta-analysis* (NMA), also known as a "multiple treatments meta-analysis" or "mixed treatment comparisons meta-analysis," compares multiple interventions at the same time by analyzing studies making different comparisons in the same analysis [19]. A key strength of NMA is the ability to combine both direct and indirect evidence in the same analysis. As a result, more useful information about what treatments work best for whom may be provided. This is especially relevant since most randomized trials compare one treatment to a control group versus other possible treatment options. For example, in a recent issue of *The Journal*, Alfonso-Cristancho et al. conducted a systematic review and NMA of randomized trials that examined the effectiveness of biologics for the management of rheumatoid arthritis [20]. Overall, and while noting the limitations of the available evidence, they found that tocilizumab was superior to cDMARDs and was as effective as other biologics for rheumatoid arthritis [20]. While the use of NMA is more recent than AD and IPD meta-analysis, it is expected that an increased number will be produced given the potential utility for clinicians and other decision-makers. To date, most NMAs have used the AD versus IPD approach, although methods for conducting a NMA based on IPD have been proposed [21]. Details regarding the conduct, reporting, and use of a NMA can be found elsewhere [22, 23].

Recently, there has been an increased interest in *non-inferiority* (NI) meta-analyses. As opposed to traditional randomized trials in which the goal is to reject the null hypothesis of no statistically significant difference between two groups, for example intervention versus control, a NI trial is designed to test whether a new intervention is no worse than the reference intervention within a certain NI margin [24]. These meta-analyses may be conducted using the AD or IPD approach. However, to the best of the authors' knowledge, no NI meta-analyses currently exist in the field of rheumatology. Additional information regarding NI meta-analyses can be found elsewhere [24].

Major elements of systematic reviews with meta-analysis

We focus the remainder of this paper on AD meta-analyses, the most common type of meta-analysis. However, much of this information can be applied to other types of meta-analyses.

An AD meta-analysis abstracts and pools summary data from individual studies. While not indigenous to AD meta-analysis, major strengths include (1) increased statistical power, (2) the potential to address uncertainty when study findings are in opposition to each other, (3) improved estimates of effect, and (4) ability to answer questions not identified in individual trials [25]. The increased power associated with an AD meta-analysis is grounded in the fact that participants are nested within studies. For example, in a recent article in *The Journal*, Mohammed et al., conducted a systematic review with AD meta-analysis of randomized controlled trials on the therapeutic effects of probiotics in participants with rheumatoid arthritis [26]. While the meta-analysis was based on only nine studies, a total of 361 participants were nested in these nine studies [26].

A systematic review with meta-analysis mirrors that of an original study: (1) Introduction, (2) Methods, (3) Results, (4) Discussion, and (5) Conclusion. Two important questions that should be addressed in the Introduction section of a manuscript are as follows: (1) What has been done before? and (2) Why is it important? Based on the authors' anecdotal experiences, one of the most overlooked aspects of meta-analyses today appears to be a lack of a thorough review of previous meta-analyses on the same topic to help justify a new or updated meta-analysis. As a result, this appears to have contributed to redundant meta-analyses that use similar or only slightly different methods to address the same or closely related research question [27]. One way to address this issue is to conduct one's own SRPSR, decide whether a new or updated meta-analysis is needed, and if so, refer to this information in the Introduction section of their own systematic review with meta-analysis. Reasons for conducting a new or updated systematic review with meta-analysis include, but are not necessarily limited to, (1) lack of an existing systematic review with meta-analysis on a topic of clinical and/or public health importance; (2) the availability of new research since the last systematic review with meta-analysis was conducted, especially if it may have an impact on previous meta-analytic findings; and (3) the use of new methods that will improve the quality of the review. Detailed guidelines for this overall process are available elsewhere [5]. Finally, the introduction section of the paper should also clearly state the purpose, i.e., research question(s) addressed, of the systematic review with meta-analysis, something that clinicians should consider with respect to potential application in their own practice.

A meta-analysis should include an a priori research plan and be described in sufficient detail so that the study can be replicated. In accordance with established guidelines [17], the methods section should usually consist of the following subsections: (1) study eligibility, (2) data sources, (3) study selection, (4) data abstraction, (5) risk of bias assessment, and (6) data synthesis. Prior to describing these areas, information should be provided that includes whether the protocol for the systematic review with meta-analysis is registered in a trial database such as PROSPERO [28] as well as whether the protocol has been published in a peer-reviewed journal. The availability of protocols for systematic reviews, with or without meta-analysis, may (1) reduce small-study effects, for example, publication bias; (2) increase transparency; and (3) avoid duplication of effort [27]. Detailed guidance for conducting and reporting protocols are provided elsewhere [29].

Study eligibility consists of what studies are to be included in a systematic review. To address this, the PICOS framework, where applicable, can be followed [17]. This includes participants/population (P), interventions (I), comparisons (C), outcomes (O), and study design/setting (S) [17]. For observational studies of etiology, consideration could be given to the PECO (population, exposure, control, outcomes) framework [18]. The inclusion of this information provides clinicians and other users of systematic reviews and meta-analyses with information regarding the utility of the systematic review with meta-analysis and subsequent application of the findings to their patients. Similar to excluded patients in an original study such as a randomized controlled trial, a list of excluded studies for the systematic review should also be provided as well as the reason(s) for exclusion. Since this is usually a long list, it is almost always preferable to include this as a supplementary file. A systematic review may include studies in any language and from both published and unpublished sources (Master's theses, dissertations, abstracts from conference proceedings, etc.). However, there is no clear consensus regarding increased bias if a systematic review is limited to English-language articles published in peer-reviewed journals [1].

Data sources consist of the methods used to search for studies. This includes such things as computer databases searched (PubMed, EMBASE, Web of Science, etc.), including search criteria for each database, cross-referencing from retrieved studies, searching clinical trials databases, hand-searching selected journals, as well as searching for grey literature from sources such as the System for Information on Grey Literature in Europe (SIGLE) and ProQuest Master's theses and dissertations databases. Information for the start and end dates of the searches should be provided along with a rationale for the starting year for searching. For example, the start date for searching for potentially eligible studies for the authors' recent systematic review with meta-analysis on exercise and anxiety in adults with arthritis and other rheumatic diseases was 1981 [30], because this was the first year that it appeared that a randomized controlled trial on exercise in adults with arthritis had been conducted. When searching for potentially eligible studies, it has been suggested that at least two electronic databases be searched [1], although it is usually wise to use more than two databases so as to not miss potentially eligible studies and thus, increase one's risk for literature search bias. The rationale for searching more than one

database is that no one database will usually include all relevant published citations. For example, Dickersin et al. reported that depending on the area or specific question, only 30–80% of all known published randomized controlled trials were identified using MEDLINE [31]. Recently, it was suggested that systematic reviewers search at least Embase, MEDLINE, Web of Science, and Google Scholar to guarantee adequate coverage [32], although searching Google Scholar may not be worth the time and effort given its lack of sensitivity and specificity [33]. In addition, it has been reported that the Scopus database provides 100% coverage of both MEDLINE and Embase [34]. Furthermore, it is important to point out that MEDLINE is nested within PubMed. Finally, the inclusion of the detailed search strategy for at least one database, for example, PubMed, should be included, often as a supplementary file. The names of the individuals responsible for conducting the searches should also be provided [17].

To avoid *study selection* bias, studies should be selected by at least two people, independent of each other. For those studies that are excluded, the reason(s) for exclusion should be reported. Following the PICOS framework, exclusions may be broadly categorized as inappropriate (1) population, (2) intervention, (3) comparison, (4) outcomes, (5) study design/setting, and (6) other. After the initial selection of studies is completed, the reviewers should then meet and discuss their selections for agreement. If agreement cannot be reached, at least one person not involved in the selection of studies should serve as an arbitrator. For all tasks, the names of the individuals responsible for each should be reported.

Data abstraction consists of the process used to code the studies. This should first include a description of how the codebooks were developed as well as a list and description of the variables that were coded. Broadly, these would include (1) study characteristics (authors, year of publication, journal, impact factor of journal, study design, etc.); (2) participant characteristics (age, gender, race/ethnicity, etc.); (3) intervention characteristics, for example, length of the study; and (4) outcome characteristics (sample sizes, means, standard deviations, etc.). Clinicians should pay particular attention to the items coded to insure that no important ones that may have affected the outcome(s) of interest are omitted. However, it is difficult to provide specific suggestions given that these variables will be highly dependent on the type of meta-analysis as well as outcome(s) of interest. Pedder et al. provide additional information regarding the abstraction of data, especially for complex meta-analyses [35].

The coding of studies should follow the same process as for the selection of studies and the process for obtaining missing data from original trial investigators should be explained, including success in retrieving missing data. While the authors are not aware of any consensus regarding the retrieval of missing data, one possible approach is to contact the corresponding author via e-mail to request such. If no response is received within one to two weeks, a second e-mail could be sent. If no response is received within the same one- to two-week period, a telephone call to request such may be appropriate. For e-mail communications, it may be beneficial to copy the other authors of the study.

An important component of any systematic review is the inclusion of some type of *risk of bias assessment* for each included study. While this is often referred to as study quality, the use of quality scales has been discouraged by some, especially with respect to the inclusion and exclusion of studies given that it is often difficult to distinguish between the quality of the reporting of a study and the quality in the conduct of a study [1]. The number of risk of bias/study quality assessment instruments available are numerous (at least 86) [36]. While varied and inconsistent, Seehra et al. reported that the Cochrane risk of bias tool was the most common tool for assessing randomized controlled trials (26.1%) while the Newcastle-Ottawa scale was the most common instrument for assessing non-randomized studies (15.3%), including case-control and cohort studies [37]. The Cochrane Collaboration recently modified their risk of bias tool for randomized controlled trials [38] and developed a tool for assessing the risk of bias in non-randomized studies that compared the health effects of two or more interventions [39]. The procedures for evaluating risk of bias are similar to those for selecting studies and abstracting data. While not without weaknesses, risk of bias results can nevertheless help clinicians and other consumers of meta-analyses with decisions regarding the strength of the included studies.

The *data synthesis* section of a systematic review can be either qualitative or quantitative, i.e., meta-analysis. With a focus on the meta-analytic approach, one of the first things that needs to be decided is the method that will be used to calculate a *common effect size* for each outcome of interest from each study so that results can be pooled. This usually includes sample sizes as well as measures of central tendency and dispersion. Whenever possible, the focus should be on calculating effect sizes using the original metric, for example body weight in kilograms, because it will be more intuitive for clinicians. However, the calculation of something like a standardized mean difference effect size (Hedge's g, Cohen's d, etc.) is usually necessary if the outcome of interest is assessed using different scales, for example, the assessment of depression using the Beck Depression Inventory, Center for Epidemiologic Studies Depression Scale, etc. [40]. Fortunately, many effect sizes can be converted between each other as well as from different tests (t tests, F-ratios, correlations, etc.) [1].

A second decision that needs to be made is the *type of model used to pool results*. Generally, there are two basic types of models, the fixed-effect model and the random-effects model. A fixed-effect model assumes that all studies share the same common effect size and that any differences in

the observed effects are the result of sampling error. In contrast, random-effects models assume that the true effect size may differ between studies. Thus, a random-effects model accounts for both sampling error and between-study variance. Since one is usually working with a universe of populations in meta-analysis, the random-effects model is usually the best choice of the two. However, since there are several random-effects models that use different methods to calculate between-study variance, it is important that authors report which model was used since they can lead to different results [41]. The most commonly used, but not necessarily the best model, is the original random-effects model of Dersimonian and Laird [42]. Its common usage is most likely the result of its longevity and availability in numerous statistical packages for meta-analysis. While random-effects models are generally recommended over the traditional fixed-effect model, the recent inverse variance heterogeneity (IVhet) as well as quality effects (QE) models have been suggested to be more robust than the traditional Dersimonian and Laird model with respect to coverage probabilities [43, 44]. However, the QE model is based on assessing a quality score for each study from a pre-existing or self-developed scale [44], something that might be difficult to do, given, as previously mentioned, the difficulty in differentiating between the quality of reporting and the quality in the conduct of a study [1].

Regardless of the model chosen, pooled results should usually be reported using point estimates and 95% confidence intervals, assuming a frequentist approach is used, as well as *z* or *t*-based alpha values. In addition, 95% prediction intervals may also be reported when models such as those based on random-effects are used [45]. Prediction intervals differ from confidence intervals in that point estimates and their confidence intervals provide an estimate of the effect size and precision while prediction intervals tell us how effects are spread around the summary effect [45]. From a clinical perspective, prediction intervals help ascertain uncertainty with respect to whether an intervention works or not. However, because of coverage problems, it has been suggested that caution be derived in drawing strong conclusions from 95% prediction intervals [46].

In meta-analysis, it is important to examine and report data on heterogeneity and inconsistency. Broadly, heterogeneity in meta-analysis refers to any type of variability between studies [1]. These could include (1) clinical factors (differences in participants, interventions, outcomes, etc.), (2) methodological features (study design, risk of bias, etc.), and (3) statistical (differences in intervention effects) [1]. Heterogeneity is typically reported using the Q statistic while inconsistency is reported using *I*-squared, an extension of Q [1]. The Q statistic is a measure of statistical significance and, given power issues, is typically reported as statistically significant if the alpha (*p*) value is less than or equal to 0.10 versus 0.05. The *I*-squared statistic is a relative measure that ranges from 0 to 100%, with higher values representing greater inconsistency. However, these estimates are not perfect and cannot explain all potential sources of heterogeneity and inconsistency [47].

A standard method of reporting the results from each study as well as the overall pooled results is through presentation of a forest plot. Figure 2 provides an example of a forest plot from the investigators' recent meta-analysis on exercise and anxiety in adults with arthritis and other rheumatic diseases [30]. If the exact same data are available from each study for calculating treatment effects, for example, sample sizes and change outcome means and standard deviations, these may also be displayed in the figure.

An important component of meta-analysis is an examination for *small-study effects* (publication bias, etc.). This should typically be examined both qualitatively, usually with some type of funnel plot, and quantitatively. Current guidelines suggest that if there are at least 10 studies, a funnel plot and Eggers's regression intercept test may be used to examine for small-study effects if the outcome of interest is continuous in nature, for example, erythrocyte sedimentation rate [48]. However, since the time of the publication of these guidelines, an alternative qualitative (Doi plot) and quantitative (LFK index) approach has been suggested to be more robust as well as allowing better coverage for meta-analyses that include less than 10 studies [49]. Figure 3 provides an example and explanation of a funnel plot, including Egger's regression intercept test, from a recent meta-analyses by the authors [30].

Because many meta-analyses include a small number of trials, typically six in Cochrane reviews [50], it is usually important to conduct *influence analysis* with each study deleted from the model once in order to examine the impact that each study has on the overall findings, something that is probably important from a clinical perspective. For example, in a recent meta-analysis by the authors, results for changes in anxiety as a result of exercise in adults with arthritis and other rheumatic diseases remained significant when each result was deleted from the model once (standardized mean difference effect size range of − 0.44 to − 0.35) [30].

In addition to influence analysis, it is usually important to conduct *cumulative meta-analysis*, typically ranked by year of publication, to examine the accumulation of results over time [51]. The inclusion of results from a cumulative meta-analysis can help clinicians and others in making more informed decisions based on past years of research as well as leading to earlier and increased use of effective interventions in practice [51]. Using this approach, results are summarized as each new study is added to the model. Thus, each horizontal line in the plot represents a pooled summary of results as each study is added versus the results of a single study in a forest plot. An example of cumulative meta-analysis from a recent meta-analysis by the authors is shown in Fig. 4 [30].

If a sufficient number of studies are available, *subgroup and/or meta-regression analyses* may be performed in order

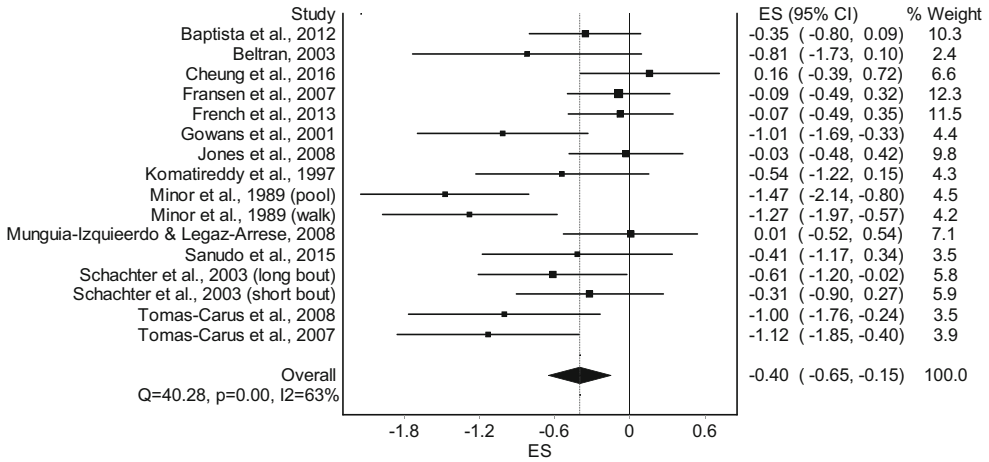

Fig. 2 Forest plot of previously published aggregate data meta-analysis addressing the effects of exercise (aerobic, strength training, or both) on anxiety in adults with arthritis and other rheumatic diseases [30]. Data are reported using the standardized mean difference effect size with results pooled using the inverse variance heterogeneity (IVhet) model. Negative values indicate improvement, i.e., reductions, in anxiety. The left and right sides of the lines represent the 95% confidence intervals for each result from each study while the black squares represent the point estimate for each result. The left and right sides of the diamond represent the pooled 95% confidence intervals while the middle of the diamond represents the overall pooled effect. The weights represent the contribution of each effect size to the overall pooled result based on the IVhet model. As can be seen by the pooled non-overlapping 95% confidence intervals, improvements were from 0.15 to 0.65 standard deviations units, values that should be given greater attention than the overall point estimate. The Cochran Q value for statistical between-study heterogeneity is statistically significant while the I^2 value is 63%, a value that was interpreted in the study as "moderate" inconsistency. *I*-squared values can range from 0 to 100% with higher values indicative of greater inconsistency. Note that two studies included more than one group, potentially impacting the independence assumption. However, results were similar when collapsed so that only one result represented each study [30]

to examine the effect of potential covariates on the outcome(s) of interest, details of which have been described elsewhere [1]. While there may be a tendency for authors to conduct such analyses only when statistically significant and/or a large amount of inconsistency exists, this is problematic given the limitations of these measures [52]. When conducting analyses such as meta-regression, the authors are not aware of any firm consensus regarding the minimum number of studies that are needed per covariate. However, as a general guideline, one may want to avoid conducting any type of meta-regression unless there are at least ten studies [1]. Also, it is important to understand that such analyses should be viewed as

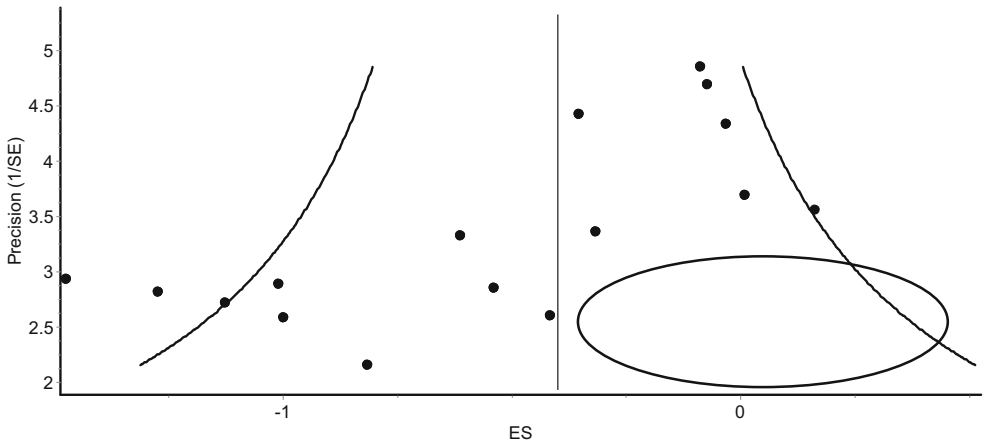

Fig. 3 Inverted funnel plot of previously published aggregate data meta-analysis addressing the effects of exercise (aerobic, strength training, or both) on anxiety in adults with arthritis and other rheumatic diseases [30]. Data on the *x*-axis represent the standardized mean difference effect size with negative values indicative of improvement, i.e., reductions, in anxiety. Data on the *y*-axis represent the precision of the estimates (inverse of the standard error). In the absence of small-study effects, effect size estimates from studies with smaller sample sizes should be spread more widely and evenly at the bottom of the funnel while studies with larger sample sizes should have effect size estimates that are closer together at the top of the funnel. The circled area with missing values in the bottom right portion of the plot is suggestive of small-study effects, possibly publication bias, i.e., avoidance of authors to submit and/or editors to publish, studies that report results that are not statistically significant. Quantitative analysis using Egger's regression intercept test supported the finding of small-study effects ($p < 0.0001$)

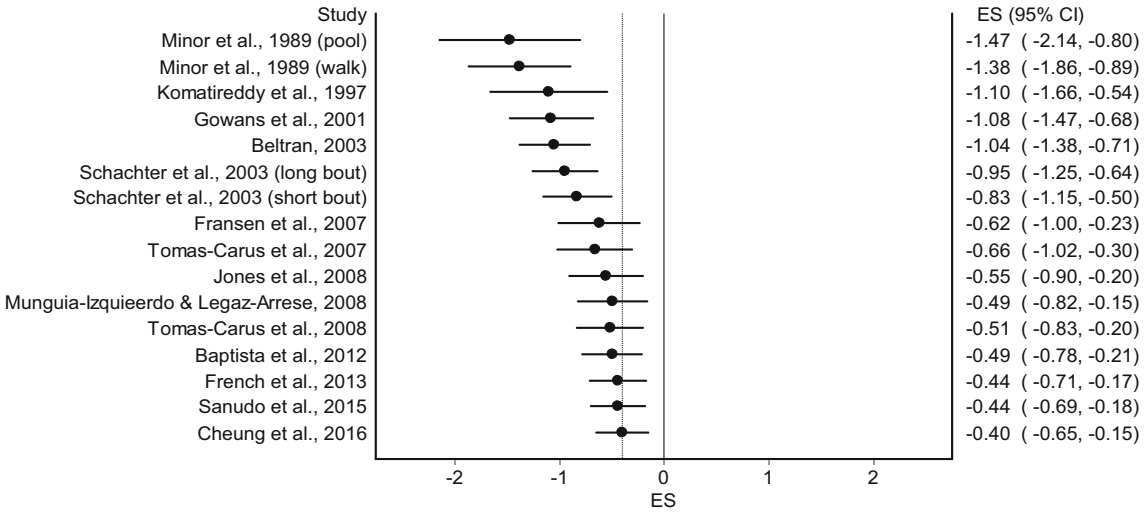

Fig. 4 Cumulative meta-analysis, ranked by year of publication, for previously published aggregate data meta-analysis addressing the effects of exercise (aerobic, strength training, or both) on anxiety in adults with arthritis and other rheumatic diseases [30]. Data are reported using the standardized mean difference effect size with results pooled using the inverse variance heterogeneity model. Negative values indicate improvement, i.e., reductions, in anxiety. The left and right sides of the lines represent the pooled 95% confidence intervals for each result from each study up to that point in time while the black circles represent the pooled point estimates for each result up to that point in time. As can be seen, results have remained statistically significant since the first study was conducted in 1989, with a trend towards smaller reductions over time, possibly reflecting better study designs with more recent studies. Note that two studies included more than one group, potentially impacting the independence assumption. However, results were similar when collapsed so that only one result represented each study

exploratory and would need to be addressed in original trials since studies are not randomly assigned to covariates in meta-analysis. Thus, they are considered to be observational in nature.

In addition to the previously mentioned information, the authors believe that it is critical that meta-analyses include metrics that are understandable and have clinical meaning. While not without limitations and potentially misleading, the use of metrics such as the *number-needed-to treat* (NNT) and *percentile improvement* based on values such as Cohen's U_3 index [53], when appropriate, should be calculated and reported. For example, the authors recently reported a standardized mean difference reduction of 0.40 in anxiety as a result of exercise in adults with arthritis and other rheumatic diseases [30]. While this means that the exercise group did 0.40 standard deviations better than the control group, it probably has little application for clinicians. In contrast, and while understanding that results were based on aggregate versus IPD, the reported NNT of 6, percentile improvement of 15.5%, and an estimated 5.3 million inactive US adults with arthritis and other rheumatic diseases improving their anxiety if they started exercising regularly, probably has much more meaning to clinicians as well as others. Whenever possible, it is also important to report both *absolute* and *relative* changes in the outcomes of interest. Furthermore, consideration should also be given to both the clinical and population health importance of the findings from a meta-analysis. For example, a 1 $kg{\cdot}m^2$ reduction in body mass index (BMI), a well-established risk factor for arthritis in adults, may not be important from a clinical, i.e., patient-level, perspective, but may have significant implications at the population level.

Finally, and possibly most relevant and time-efficient for clinicians, the *strength of the evidence* should usually be assessed and reported. One common way to address this is by use of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) instrument, details of which are described elsewhere [54]. Briefly, GRADE, a subjective instrument, assesses the strength of evidence for a specific outcome with respect to (1) risk of bias, (2) imprecision, (3) inconsistency, (4) indirectness, and (5) publication bias [54]. Within these categories, the quality of the evidence can be rated down by one to two levels for any one of these items. The quality of the evidence can also be increased by one or two levels if there is a large magnitude of effect or one level if either a dose-response gradient is found or all plausible confounding would decrease the effect or increase the effect if indeed no effect was found [54]. Risk of bias has to do with study limitations and consists of factors such as lack of allocation concealment and blinding, incomplete accounting of participants and outcome events, selective outcome reporting, as well as any other limitations that reviewers think may impact the outcome [54]. Imprecision has to do with uncertainty about the results, such as a wide confidence interval around the estimate of effect while inconsistency refers to unexplained heterogeneity in results [54]. Indirectness consists of evaluating results based on whether the included studies directly compare the interventions and populations that one is interested in as well as assessing outcomes deemed important

Table 1 Five basic questions for clinicians to ask when applying meta-analytic results to practice

1. Are the research questions addressed applicable to my patients?
2. Is the population of patients studied similar to those that I see in my practice?
3. How does this treatment compare to other alternatives?
4. Do the benefits of the treatment outweigh the harms?
5. What is the strength of the evidence?

by participants, for example physical function in those with arthritis [54]. Finally, publication bias has to do with the selective publication of studies in which benefits are overestimated and harms are underestimated [54]. The overall certainty of the evidence is then rated by the authors as (1) very low, (2) low, (3) moderate, or (4) high [54]. For example, a recent meta-analysis in *The Journal* on interventions for reducing anxiety and depression in adults with rheumatoid arthritis reported the overall strength of evidence, based on GRADE, to be "moderate," meaning that the authors believed that the true effect was probably close to the estimated effect [55].

Concluding remarks

The purpose of this article was to provide a gentle introduction to clinicians with respect to systematic reviews and meta-analysis, including the potential clinical utility of such. In the spirit of that goal, Table 1 provides a list of five general questions for clinicians to consider with respect to applying meta-analytic results to their own clinical practice. It is the author's hopes that the information provided in this article will aid clinicians in making more informed decisions regarding the usability of meta-analysis in their own daily practice, with the ultimate goal of improving patient care.

Compliance with ethical standards

Disclosures None.

References

1. Higgins JPT, Green S (eds) (2011) Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration
2. Pham MT, Rajic A, Greig JD, Sargeant JM, Papadopoulos A, McEwen SA (2014) A scoping review of scoping reviews: advancing the approach and enhancing the consistency. Res Synth Methods 5(4):371–385
3. Misra DP, Sharma A, Kadhiravan T, Negi VS (2017) A scoping review of the use of non-biologic disease modifying anti-rheumatic drugs in the management of large vessel vasculitis. Autoimmun Rev 16(2):179–191
4. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L, Hempel S, Akl EA, Chang C, McGowan J, Stewart L, Hartling L, Aldcroft A, Wilson MG, Garritty C, Lewin S, Godfrey CM, Macdonald MT, Langlois EV, Soares-Weiser K, Moriarty J, Clifford T, Tunçalp Ö, Straus SE (2018) PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation: the PRISMA-ScR statement. Ann Intern Med 169(7):467–473
5. Kelley GA, Kelley KS (2018) Systematic reviews and cancer research: a suggested stepwise approach. BMC Cancer 18(246):9
6. McKenzie JE, Brennan SE (2017) Overviews of systematic reviews: great promise, greater challenge. Syst Rev 6(1):185
7. Christie A, Jamtvedt G, Dahm KT, Moe RH, Haavardsholm EA, Hagen KB (2007) Effectiveness of nonpharmacological and non-surgical interventions for patients with rheumatoid arthritis: an overview of systematic reviews. Phys Ther 87(12):1697–1715
8. Fusar-Poli P, Radua J (2018) Ten simple rules for conducting umbrella reviews. Evid Based Mental Health 21(3):95–100
9. Tierney JF, Vale C, Riley R, Smith CT, Stewart L, Clarke M, Rovers M (2015) Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use. PLoS Med 12(7):e1001855
10. Riley RD, Simmonds MC, Look MP (2007) Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol 60(5):431–439
11. Kelley GA, Kelley KS (2016) Retrieval of individual participant data for exercise meta-analyses may not be worth the time and effort. Biomed Res Int 2016(5059041):1–5
12. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, Boutitie F (2008) Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Stat Med 27(11):1870–1893
13. Steinberg KK, Smith SJ, Stroup DF, Olkin I, Lee NC, Williamson GD, Thacker SB (1997) Comparison of effect size estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. Am J Epidemiol 145:917–925
14. Cooper H, Patall EA (2009) The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. Psychol Methods 14(2):165–176
15. Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, Tierney JF (2015) Preferred reporting items for systematic review and meta-analyses of individual participant data: the PRISMA-IPD statement. JAMA 313(16):1657–1665
16. Higgins J, Green S (2008) Cochrane handbook for systematic reviews of interventions (version 5.0.1)
17. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 151(4):W65–W94
18. Dekkers OM, Vandenbroucke JP, Cevallos M, Renehan AG, Altman DG, Egger M (2019) COSMOS-E: guidance on conducting systematic reviews and meta-analyses of observational studies of etiology. PLoS Med 16(2):e1002742
19. Petticrew M, Rehfuess E, Noyes J, Higgins JPT, Mayhew A, Pantoja T, Shemilt I, Sowden A (2013) Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. J Clin Epidemiol 66(11):1230–1243
20. Alfonso-Cristancho R, Armstrong N, Arjunji R, Riemsma R, Worthy G, Ganguly R, Kleijnen J (2017) Comparative effectiveness of biologics for the management of rheumatoid arthritis: systematic review and network meta-analysis. Clin Rheumatol 36(1):25–34

21. Riley RD, Jackson D, Salanti G, Burke DL, Price M, Kirkham J, White IR (2017) Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. Br Med J 358
22. Mills EJ, Ioannidis JP, Thorlund K, Schunemann HJ, Puhan MA, Guyatt GH (2012) How to use an article reporting a multiple treatment comparison meta-analysis. JAMA 308(12):1246–1253
23. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, Ioannidis JP, Straus S, Thorlund K, Jansen JP, Mulrow C, Catala-Lopez F, Gotzsche PC, Dickersin K, Boutron I, Altman DG, Moher D (2015) The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med 162(11):777–784
24. Brittain EH, Fay MP, Follmann DA (2012) A valid formulation of the analysis of noninferiority trials under random effects meta-analysis. Biostatistics 13(4):637–649
25. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC (1987) Meta-analysis of randomized controlled trials. N Engl J Med 316:450–455
26. Mohammed AT, Khattab M, Ahmed AM, Turk T, Sakr N, Khalil AM, Abdelhalim M, Sawaf B, Hirayama K, Huy NT (2017) The therapeutic effect of probiotics on rheumatoid arthritis: a systematic review and meta-analysis of randomized control trials. Clin Rheumatol 36(12):2697–2707
27. Ioannidis JPA (2016) The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q 94(5):485–514
28. Page MJ, Shamseer L, Tricco AC (2018) Registration of systematic reviews in PROSPERO: 30,000 records and counting. Syst Rev 7(1):32
29. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA (2015) Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. Br Med J 349:g7647
30. Kelley GA, Kelley KS, Callahan LF (2018) Community-deliverable exercise and anxiety in adults with arthritis and other rheumatic diseases: a systematic review with meta-analysis of randomised controlled trials. BMJ Open 8(2):18
31. Dickersin K, Scherer R, Lefebvre C (1994) Identifying relevant studies for systematic reviews. Br Med J 309:1286–1291
32. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH (2017) Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev 6:245
33. Vine R (2006) Google Scholar. J Med Libr Assoc 94(1):97–99
34. Burnham JF (2006) Scopus database: a review. Biomed Digit Libr 3(1):1–8
35. Pedder H, Sarri G, Keeney E, Nunes V, Dias S (2016) Data extraction for complex meta-analysis (DECiMAL) guide. Syst Rev 5(1): 212
36. Sanderson S, Tatt ID, Higgins JP (2007) Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 36(3):666–676
37. Seehra J, Pandis N, Koletsi D, Fleming PS (2016) Use of quality assessment tools in systematic reviews was varied and inconsistent. J Clin Epidemiol 69:179–184
38. Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, Reeves B, Eldridge S (2016) A revised tool for assessing risk of bias in randomized trials. Cochrane Database Syst Rev 10(Supplement 1)
39. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan A-W, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JP (2016) ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 355
40. Kelley GA, Kelley KS, Hootman JM (2015) Effects of exercise on depression in adults with arthritis: a systematic review with meta-analysis of randomized controlled trials. Arthritis Res Ther 17:21
41. Zeng D, Lin DY (2015) On random-effects meta-analysis. Biometrika 102(2):281–294
42. Dersimonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 7:177–188
43. Doi SA, Barendregt JJ, Khan S, Thalib L, Williams GM (2015) Advances in the meta-analysis of heterogeneous clinical trials I: the inverse variance heterogeneity model. Contemp Clin Trials 45(Pt A):130–138
44. Doi SA, Barendregt JJ, Khan S, Thalib L, Williams GM (2015) Advances in the meta-analysis of heterogeneous clinical trials II: the quality effects model. Contemp Clin Trials 45(Pt A):123–129
45. Higgins JP, Thompson SG, Spiegelhalter DJ (2009) A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A 172(1):137–159
46. Partlett C, Riley RD (2017) Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following REML estimation. Stat Med 36(2):301–317
47. Ioannidis JP, Patsopoulos NA, Evangelou E (2007) Uncertainty in heterogeneity estimates in meta-analyses. BMJ 335(7626):914–916
48. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, Carpenter J, Rucker G, Harbord RM, Schmid CH, Tetzlaff J, Deeks JJ, Peters J, Macaskill P, Schwarzer G, Duval S, Altman DG, Moher D, Higgins JP (2011) Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. Br Med J 343:d4002
49. Furuya-Kanamori L, Barendregt JJ, Doi SAR (2018) A new improved graphical and quantitative method for detecting bias in meta-analysis. Int J Evid Based Healthc 16(4):195–203
50. Mallett S, Clarke M (2002) The typical Cochrane review. How many trials? How many participants? Int J Technol Assess Health Care 18(4):820–823
51. Clarke M, Brice A, Chalmers I (2014) Accumulating research: a systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. PLoS One 9(7):e102670
52. Higgins J, Thompson S, Deeks J, Altman D (2002) Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. J Health Serv Res Policy 7(1):51–61
53. Cohen J (1988) Statistical power analysis for the behavioral sciences. Academic Press, New York
54. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ (2008) GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. Br Med J 336(7650):924–926
55. Fiest KM, Hitchon CA, Bernstein CN, Peschken CA, Walker JR, Graff LA, Zarychanski R, Abou-Setta A, Patten SB, Sareen J, Bolton J, Marrie RA, Burden CTDt (2017) Managing the effects of psychiatric comorbidity in chronic immunoinflammatory D (2017) systematic review and meta-analysis of interventions for depression and anxiety in persons with rheumatoid arthritis. J Clin Rheumatol 23(8):425–434