Systematic Reviews and Meta-Analyses of Oncology Studies

18

Allan A. Lima Pereira and Andre Deeke Sasse

18.1 Introduction

The current volume of research articles published every year is in continuous growth and it has become virtually impossible for physicians, even when they are focused on more specific fields, to keep up with the enormous amount of research data. The major reason for conducting a review is that large quantities of information must be simplified into palatable parts for understanding.

There are different types of reviews. Not all review articles are systematic reviews and not all systematic reviews are followed by a meta-analysis. Reviews that do not use planned scientific methods to search, collect, and summarize information are not systematic reviews. They usually are traditional narrative reviews, where there are no clearly specified methods of identifying, selecting, and validating information included from multiple studies. Once systematic reviews have been performed, only a subset of them will include statistical methods to quantify and combine the results from independent studies, which we call meta-analysis.

Commonly in oncology, there are controversies about the real value of interventions. It is, therefore, important to recognize potential biases and also to establish as accurately as possible the actual differences between the strategies being evaluated. Summarizing the evidence facilitates the interpretation of the results, and makes it possible to identify whether the claimed statistically significant benefits are also

A. A. Lima Pereira, M.D., Ph.D.
Department of Gastrointestinal Medical Oncology,
University of Texas – M.D. Anderson Cancer Center, Houston, TX, USA

A. D. Sasse, M.D., Ph.D. (✉)
Department of Internal Medicine, Faculty of Medical Sciences,
University of Campinas (UNICAMP), Campinas, SP, Brazil
e-mail: sasse@cevon.com.br

R. L. C. Araújo, R. P. Riechelmann (eds.), *Methods and Biostatistics in Oncology*,
https://doi.org/10.1007/978-3-319-71324-3_18

clinically relevant. For this reason, systematic reviews are needed to refine the unmanageable amounts of information found in electronic databases, separating the insignificant, unsound, or redundant deadwood in the literature from the studies that are worthy of reflection [1], and then using the processed information for different purposes, such as to:

- Make recommendations for clinical practice and guidelines
- Establish the state of existing knowledge (useful when applying for grants)
- Clarify conflicting data from different studies
- Highlight areas where further original research are required.

Also, many times, a meta-analysis can add better quality evidence to the current medical literature. For instance, after pooling together many underpowered negative studies, a meta-analysis can finally give us the answer that each study alone was unable to provide. However, if not done properly, a meta-analysis can lead to bias (metabias). In addition, systematic reviews have become impressively more common [2]. Therefore, it is crucial that physicians become familiar with interpreting this kind of work; the best way to do this is to gain understanding of the key points needed to perform such work.

18.2 How to Plan a Systematic Review

The first step in performing a systematic review is to define the research question. However, to avoid waste of time or duplication of efforts, it is important to search for published and ongoing systematic reviews which might have already answered the same question or are aiming to do so. This search can be made in specific databases, such as the Cochrane Library (http://www.cochranelibrary.com) and PROSPERO (www.crd.york.ac.uk/PROSPERO). General databases (e.g., MEDLINE and EMBASE) should also be searched.

After the research question has been decided and the need for a new review has been confirmed, a protocol should be registered in public databases (such as Cochrane and PROSPERO). A written protocol defines the study methodology and sets the inclusion/exclusion criteria for trials, literature searches, data extraction and management, assessment of the methodological quality of individual studies, and data synthesis. As for any clinical study, the systematic review protocol must be designed a priori. Although the majority of oncology medical journals do not require an a priori registered protocol, we believe this is necessary to minimize the risk of systematic errors or biases being introduced by decisions that are influenced by the findings.

Ideally, a systematic review and its protocol are planned and conducted by a team with multiple skills. A team leader should coordinate and write the final report. A medical oncologist with clinical practice is needed to clarify issues related to the chosen topic. Reviewers are required to screen abstracts, read the full text, and extract the data. A statistician can assist with data analysis. Frequently, researchers

accumulate different functions, but a well-planned team helps to reduce the risk of errors; a team of at least three people is needed.

18.2.1 Framing the Question

As mentioned above, the beginning of a systematic review occurs through building a good clinical question. A well-formulated question usually has four parts: the population, the intervention; the comparison intervention; and the outcome. This question structure is known by the acronym PICO (Problem/Patient/Population, Intervention/Indicator, Comparison, Outcome). The PICO framework helps to identify key concepts of the question, and should be sufficiently broad to allow examination of variation in the study factor (e.g., intensity or dose regimen) and across populations. An example of a good and straightforward clinical question using the PICO framework can be found in a published systematic review [3] and is detailed below:

– P: metastatic colorectal cancer patients receiving first-line systemic palliative treatment
– I: complete stop of treatment
– C: continuous treatment until disease progression
– O: overall survival.

Therefore, the question is: "Does complete stop of treatment in the first-line palliative setting of metastatic colorectal cancer patients impact overall survival?" Note this final question allows the inclusion of different regimens, durations, and intensities, and makes it possible to evaluate only the strategy of concern. The decision of how broad or narrow a clinical question to use is based on clinical judgment. A "narrower" question may not be clinically useful and can result in false or biased conclusions. On the other hand, broad questions may pool together studies too different to be combined ("apples with oranges") and make the search process more difficult and time-consuming.

Framing the question is not only the first step of a systematic review. It is also the most important, since it will have a direct impact on the inclusion and exclusion criteria used to select studies, the development of the search strategy, and the main data to be abstracted.

18.2.2 Searching the Evidence

It is easy to find a few relevant articles by a straightforward literature search, but the process becomes progressively more challenging as we try to find more "hidden" trials.

Systematic reviews of interventions require a thorough, objective, and reproducible search of a range of sources to identify as many relevant studies as possible. A search of PubMed/MEDLINE alone is not considered adequate. It is known that

only 30% to 80% of all known published randomized trials are identifiable using MEDLINE [4]. In the field of oncology, it is critical to search electronic databases such as MEDLINE and EMBASE, but also databases from clinical trials, and summaries as the Cochrane Library. However, searching the LILACS database is irrelevant in systematic reviews in oncology [5].

It is essential to define in advance structured and highly sensitivity search strategies for the identification of trials in each database. These strategies should be described later in the formal article, to allow reproducibility. There are no magic formulae to make all of the process easy, but there are some standard tactics which could be helpful.

A central tactic for a good literature search in the electronic databases is to take a systematic approach to breaking down the review question into components, which can be combined using "AND" and "OR" terms. Using the example above, in the review evaluating "Does complete stop of treatment in the first-line palliative setting of metastatic colorectal cancer patients impact overall survival?", the key components:

- (colorectal neoplasms AND maintenance chemotherapy) represent the overlap between these two terms and retrieve only articles that use both terms. A PubMed search using these terms retrieved 279 articles (at the time of all searches, in April, 2017: new citations are added to the PubMed database regularly).
- (colorectal neoplasms AND (maintenance chemotherapy OR intermittent chemotherapy)) represents a broader search, which includes other possible terms in the articles that can describe the strategies. A PubMed search using these terms retrieved 513 articles.
- (colorectal neoplasms AND maintenance chemotherapy AND intermittent chemotherapy) represents the small set where all three terms overlap. A PubMed search using these terms retrieved only 13 articles.
- (colorectal neoplasms AND (maintenance chemotherapy OR intermittent chemotherapy) AND random*) combines the term random*, which is the shorthand for words beginning with random, e.g., randomized, randomization, randomly. A PubMed search using these terms retrieved 20 articles.

Although the overlap of all three terms will usually have the best concentration of relevant articles, this strategy will probably miss many relevant studies. The ideal search strategy combines precision with sensitivity.

Usually, the initial strategy will inevitably miss useful terms, and the search process will need to be repeated and refined. However, the results of initial searches are used to retrieve the initial relevant papers, which can be used in two ways to identify missed trials:

- The bibliographies of the found articles can be checked for articles missed by the initial search;
- A citation search, using the Science Citation Index, can be conducted to identify papers that have cited the identified studies, some of which may report subsequent primary research.

The missed paper can provide clues on how the search may be broadened to capture further papers, sometimes using other keywords. The whole process may then be repeated using the new keywords identified.

It is important to remember that studies are conducted in all parts of the world, and may be published in different languages. Ideally, a systematic review should include all relevant studies, irrespective of the publication language. Including articles written only in English would lead to greater biases, as positive studies conducted in countries where English is not the state language are more likely than negative ones to be submitted to an English-language journal. This increases the usual publication bias with an additional "tower of Babel" bias.

Having a reviewer who has good experience with databases is crucial for building an efficient literature search. But the use of multiple strategies is important to track down all relevant articles. As the whole process is complex and has a high risk of loss due to fatigue, it is fundamental that the literature searches should be done by two researchers, independently.

Duplicate publications and reports should be handled with caution. Systematic reviews have studies as the primary units of interest and analyses. However, a single study may have more than one report about the results. Each report should be analyzed and each may contribute useful information for the review. Thus, no publication should be discarded solely because of duplication. However, only the most complete or most recent data should be used in the final analyses, and the duplicates should be highlighted in the flowchart of paper selection.

18.2.2.1 Publication Bias

As one could expect, it has been demonstrated that statistically significant findings have a higher likelihood of being reported than non-significant ones [6–9]. Because of such publication bias, potentially relevant studies could be missing from a meta-analysis.

There are different ways to assess whether publication bias is present in a meta-analysis. The most commonly used methods are based on funnel plot asymmetry [10–12] (Fig. 18.1). In a funnel plot, each study's treatment effect (shown on the *x*-axis) is plotted against a measure of that study's size or precision, usually using the standard error of the treatment effect on a reverse scale (shown on the *y*-axis). The name "funnel plot" comes from the fact that the accuracy of the estimate of the effect increases as the sample size increases. Thus, in the absence of publication bias, the studies will be dispersed in a *symmetrically* inverted funnel format. Studies with smaller sample sizes, which lack power and precision, will usually be spread at the bottom. As larger studies are published, the effect estimate tends to remain the same, due to the increasing accuracy, configuring the vertex of the funnel. Nevertheless, there are points of criticism about this method. First, some authors have argued that the visual interpretation of funnel plots is too subjective to be used [13]. Second, other explanations for asymmetry include heterogeneity and methodological anomalies. Finally, as Sterne et al. [14] suggest, the number of studies required to test selection bias by funnel plot should be ten or larger.

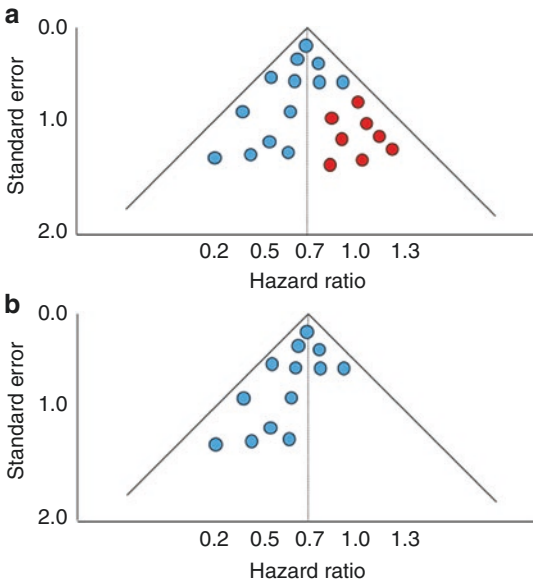

Fig. 18.1 Two hypothetical scatter plots of measure of study size vs. measure of treatment effect, known as *funnel plots*. Each dot represents a study. (**a**) Symmetrical funnel plot, suggesting absence of publication bias. (**b**) Asymmetrical funnel plot, with an apparent absence of studies with non-significant hazard ratios (HR ~ 1.0). Adapted from Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol. 2001;54 (10):1046–55

18.2.2.2 Gray Literature and Hand-Searching

To minimize the risk of selection bias, it is crucial to find all important data, and also to critically evaluate all existing pieces of evidence, including gray literature, which can be defined as unpublished studies or studies that are not commercially published and, therefore, are not indexed in the relevant databases [15]. In oncology, the more common sources of gray literature are regulatory information (The United States Food and Drug Administration [FDA] and the European Medicines Agency [EMA]), trial registers (clinicaltrials.gov), and conference abstracts. The Scopus and EMBASE databases usually provide unpublished works presented at the main oncology conferences. Other examples of gray literature are book chapters, pharmaceutical company data, letters, dissertations, and theses.

It has been shown that published papers, compared with gray literature, yield significantly larger estimates of the intervention effect [15–18]. Therefore, many argue in favor of including studies from the gray literature in order to more precisely estimate the intervention effect. On the other hand, unpublished studies and studies published in the gray literature lack peer review and might be incomplete, which raises concerns regarding their methodological quality, leading others to question whether they must be included in a meta-analysis. Despite the controversy, the acceptance of gray literature in systematic reviews by researchers and editors is increasing [19, 20] and guidelines for reporting systematic reviews, such as PRISMA [21, 22], AMSTAR [23], and Cochrane [24] recommend that researchers should identify and include all reports, gray and published, that meet the predefined inclusion criteria.

Following the same reasoning as that for searching gray literature, it is suggested that a "hand-search" be performed of the references in the included studies or those

in previous reviews. This action can be useful in identifying eligible articles that may not have been retrieved by the search strategy.

18.3 Dealing with Data

18.3.1 Extracting the Data

For most systematic reviews, data collection forms are essential for dealing with published or presented studies. The data collection form is not reported itself, but it is a bridge between what is reported by the original researchers and what is ultimately reported by the reviewers. A good form should include details about the identification of trials, the inclusion/exclusion criteria, risk of bias, methodological aspects of trials, and, finally, data for inclusion in the analysis. Because each systematic review is different, data collection forms will vary across reviews.

It is highly recommended that more than one reviewer extract data from each report, to minimize errors and reduce potential biases that could be introduced by review authors. It has already been shown that, although single data extraction requires less time and fewer human resources, it generates more errors [25]. Special attention should be given to endpoints involving subjective interpretation. Disagreements between reviewers should be recorded and described in the final publication.

When studies are reported in more than one publication or presentation, the data should be extracted from each report separately, and afterward the reviewers should combine information across multiple data collection forms.

Frequently, overall survival (OS) and other time-to-event outcomes (such as progression-free survival or disease-free survival) are evaluated in oncological systematic reviews. These endpoints are best evaluated using the hazard ratio (HR) [26], which is presented with the respective confidence interval (CI). Dichotomous data (such as response rates and adverse events) are usually analyzed using the odds ratio (OR). More rarely the risk ratio (RR) can also be presented.

Sometimes HRs are not presented for OS analyses. However, in almost all cases it is possible to calculate estimates by transcribing the survival curves presented or by using other original data with a spreadsheet developed by Tierney et al. and available online [27]. Continuous outcomes, with mean values and standard variation, are not frequent in oncology trials.

18.3.2 Assessing the Risk of Bias

It is important to understand that, whereas in a clinical study the individual is usually a patient, in a systematic review with/without meta-analysis the individual is a study. Therefore, one pitfall of systematic reviews and meta-analysis is that they are subject to the validity and quality of the studies included. In fact, one can apply a common concept of computer science called “garbage in, garbage out”,

where the quality of the output (results from a meta-analysis) is determined by the quality of the input (included studies). Therefore, all studies that meet the eligibility criteria for the systematic review must have their methodological quality assessed on an individual basis. Problems with the design and execution of individual trials raise questions about the internal and external validity of their findings and there is evidence to conclude that biases are introduced into the results of a meta-analysis when the methodological quality of the included studies is inadequate (even when they are randomized controlled trials) [28]. "Study quality" and "risk of bias", will be used here as synonymous, although the Cochrane Collaboration favors "risk of bias" instead of "quality", as "an emphasis on risk of bias overcomes ambiguity between the quality of reporting and the quality of the underlying research" [24].

The issues of quality assessment are not always related to the design of the trial. Often the trials are just poorly described. In fact, whenever we face an article, we are almost never able to find out how well the study was performed. The only information available for making a judgment regarding a study's risk of bias is the way that it was reported. In other words, we are only able to evaluate how well it was reported. For instance, we are usually not able to evaluate the quality of study procedures, protocol violations, or whether there was any data fabrication or falsification, simply because this information is not usually reported.

Currently, a large number of tools are available for assessing the methodological quality of studies (e.g., the Cochrane tool [29], Jadad [30], and Delphi [31], among others). A meta-analysis may include only high-quality trials; alternatively, a sensitivity analysis (see Sect. 18.4.4) can be done according to the quality of the trials. Each tool has its own instructions, and a detailed description of each one is beyond the scope of this chapter. Items described under the following headings are general concepts of the key methodological subjects often assessed by these tools (discussed more deeply in Chap. 10).

18.3.2.1 Randomization and Allocation Concealment

The included article should report whether randomization was done, and if so, the method used. Random numbers tables, computer random number generators, and stratified or block randomization or minimization are considered to be methods with a low risk of bias. The use of date of birth or date of visit/admission (e.g., even or odd dates) is at high risk of bias. Allocation concealment is responsible for maintaining the effect of randomization in preventing selection bias. The article should report the allocation concealment method. Methods that adequately prevent investigators from predicting the type of group to which the patients were allocated, such as central allocation (e.g., phone, web, or pharmacy), are considered as having a low risk of bias. Trials in which randomization is inadequately concealed are more likely to show a beneficial effect of the intervention [32]. After analyzing 102 meta-analyses that examined 804 trials, of which 272 (34%) had adequate allocation concealment, Wood et al. showed that trials with unclear or inadequate allocation concealment tended to show a more favorable effect of the experimental treatment [33].

18.3.2.2 Blinding/Masking

Low risk of bias means that it is unlikely that the blinding could have been broken or, in the case of an open trial, that the outcome would not be influenced by an inclusion of blinding. Although the lack of blinding has little or no effect on objective outcomes (such as death/OS), it usually yields exaggerated treatment effect estimates for subjective outcomes (such as pain levels). Wood et al. also showed, based on 76 meta-analyses examining 746 trials, of which 432 (58%) were blinded, that intervention effects can be exaggerated by 7% in non-blinded compared with blinded trials [33].

18.3.2.3 Losses to Follow-Up/Exclusions/Missing Data

Incomplete outcome data are due to patient dropouts or exclusions and there are a number of reasons why they occur. It is assumed that the higher the proportion of missing outcomes, or the larger the difference in proportions between the groups, the higher is the risk of bias. Also, there is the theoretical risk that investigators could have excluded patients to favor the experimental intervention. In addition, all randomized patients must be included in the analysis ("intention-to-treat analysis"), which means that a patient who did not receive the intervention, as mandated by protocol, for any reason should not be excluded from the final analysis.

18.3.3 Qualitative Analysis

Although not all systematic reviews have a meta-analysis, they do all have a qualitative analysis, which is presented in the "Results" section of a systematic review. A qualitative analysis usually begins by describing the search process, illustrated by a flow chart, specifying the databases and the number of records retrieved, and giving reasons why studies were excluded. This description gives the reader an idea of the comprehensiveness of the search strategy and increases the internal validity of the review.

It is also during the qualitative analysis that the authors highlight the clinical and methodological characteristics of the included studies, including their size, design, inclusion/exclusion of important subgroups, strengths, and limitations, and the relationships between the study characteristics and the authors' reported findings. All data of interest extracted from each included study, regardless of the number of articles eligible, should be compiled in the form of Tables, making it easier for the reader to have an overview of the studies' main characteristics, including some kind of clinical heterogeneity among the studies.

18.4 Meta-Analysis: Summarizing Results Across Studies

They may seem complex, but all commonly used methods for meta-analyses follow some common principles. Meta-analysis is basically a two-stage process. In the first stage, a summary statistic is calculated for each trial, to describe the observed intervention effect, which is based on the type of variable (Table 18.1). In the second

Table 18.1 Types of variables and their corresponding measures of effect

Type of variable	Effect measures
Dichotomous	Risk ratio (relative risk)
	Odds ratio
	Risk difference
Continuous	Mean difference (difference in means)
	Standardized mean difference
Ordinal	Proportional odds ratios
	Same as dichotomous[a]
	Same as continuous[a]
Time-to-event	Hazard ratio

[a]In practice, longer ordinal scales are often analyzed as continuous data and shorter ordinal scales are often made into dichotomous data by combining adjacent categories together

stage, a pooled intervention effect estimate is calculated as a weighted average of the intervention effects estimated in the individual trials. The weights of each study are chosen to reflect the amount of information that each trial contains, correlated with the sample sizes and dispersion of data; the weights are based on the analysis model (fixed-effect model vs. random-effects model) and the statistical method chosen.

18.4.1 Fixed-Effect Model Vs. Random-Effects Model

The fixed-effect model is based on the mathematical assumption that there is a single common treatment effect (one true effect size) across the studies, and the differences among the effect estimates of each study are attributed merely to chance or type-II errors. If all studies were infinitely large, they would share the same estimates of effect. Therefore, if you consider that all included studies are functionally identical, and have very similar populations and the same experimental and control interventions, a fixed-effect model may be applied. This model will compute the common effect size for this specific population in a more precise manner than the random-effects model, but you should not extrapolate your findings to other populations. This is a rare situation in oncology.

In contrast to the fixed-effect model, the random-effects model assumes that the true effect of the intervention might be different across the studies. This model allows that the true effect size may differ from study to study by chance. This is the reason why the word "effect" is singular in "fixed-effect model" (one true effect) and plural in "random-effects model" (multiple true effects). The random-effects method will usually provide an estimate of the effect with less precision (i.e., with a wider CI), which can be considered a more conservative approach and is indicated in the vast majority of meta-analyses. A recent review of systematic reviews in oncology showed that the random-effects model was underused [34].

Statistically speaking, when using a fixed-effect model, you are pooling together the observed effects from each study (the data you extract from articles) and combining them to make your best guess of what the true common effect they all share really is. Again, if each study was perfect and infinitely large, the observed effects

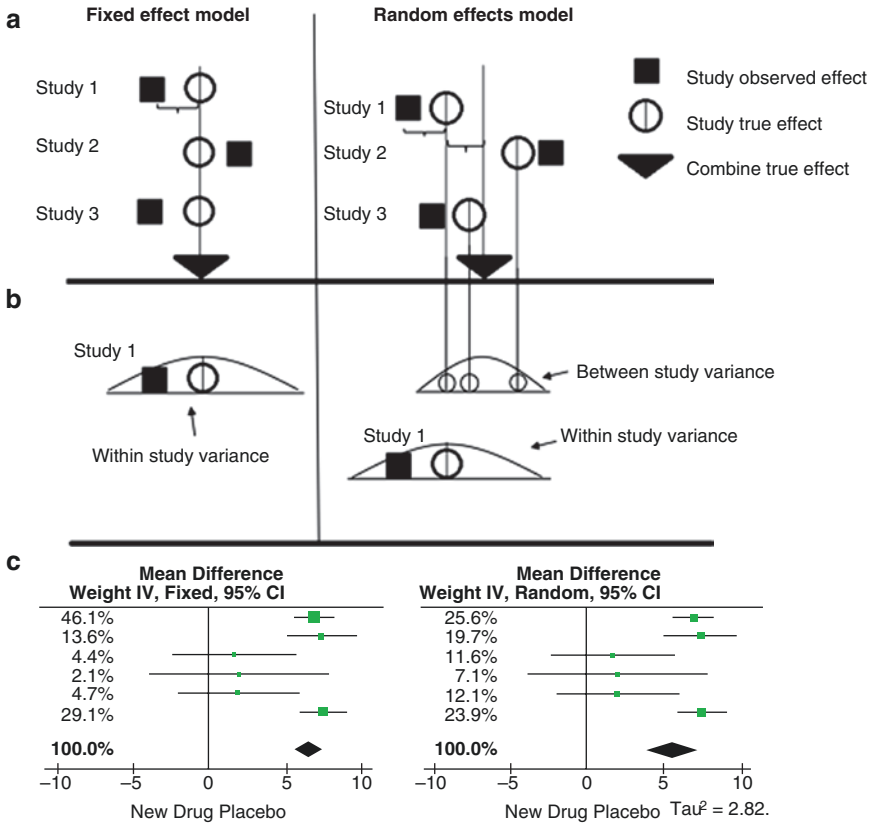

Fig. 18.2 Differences between fixed- and random-effects models. (**a**) The difference between the observed effect (filled square) and the combined true effect has one component in the fixed model and two in the random model. (**b**) This fact leads to one source of variance in the fixed-effect model, while the random-effects model has two sources. (**c**) Example of fixed-effect and random-effects meta-analyses with the same studies. The impact of the method chosen on the weight of each study results in significant differences in the sizes of the squares and the width of the diamonds. Adapted from Borenstein M et al. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010;1 (2):97–111

of each study would be the same and equal to the true effect (Fig. 18.2a). The difference between the observed effects in each study from the one common true effect they all share is due only to random errors inherent to each study. Therefore, the fixed-effect model has only one source of variance: the within-study variance. In contrast, in a random-effects model, there are two sources of variance: the within-study variance and the between-study variance. The latter is represented by $\tau 2$ (Tau-square). The weight each study receives is (often) the inverse of variance (see Sect. 18.4.2.1). However, while in the fixed-effect model the variance has one component, the random-effects model has two [35]. Therefore, statistically, the only difference between the fixed and random models is how your software weights each

study. The weight equals the inverse of the variance in both models, but the variance is further modified by the between-study variance in the random-effects model by using $\tau 2$ (Tau-square). Note that, as the meta-analysis shown references in Fig. 18.2c used the random-effects model, the Tau-square was shown (it would be absent in the case of a fixed-effect model).

18.4.2 Statistical Methods

A number of available statistical methods are used to weight effect estimates among the studies included in a review and to pool them together. Three of the most common methods are are outlined below.

18.4.2.1 Generic Inverse-Variance Method

The generic inverse-variance method is one with high applicability because it combines any effect estimates that have the standard error reported. This method can be used to combine dichotomous or continuous data and for fixed- and random-effects models.

Mathematically, variance is the square of the standard error. In turn, standard error describes the extent to which the estimate may be wrong owing to random error. The bigger the sample size of a study, the smaller are both the variance and the standard error. The inverse-variance method assumes that the variance is inversely proportional to the importance of the study; that is, the lower the variance, the more weight will be attributed to this study.

18.4.2.2 Mantel-Haenszel Method

When the data of the studies are scarce in terms of events and/or the studies have small sample sizes, estimates of the standard errors of the effect by inverse variance methods may be poor. In such situations, the Mantel-Haenzel method is preferable, since it uses a different model of weight assignment from that used for the inverse of the variance. This method is used only for dichotomous data, but can be used for both fixed- and random-effects models.

18.4.2.3 Peto Odds Ratio Method

This method is used only for dichotomous data that used the OR as an effect measure and only for the fixed-effect model. It is an alternative to the Mantel-Haenszel method, and is preferable when the two treatment arms have roughly the same number of participants and the treatment effect is small (ORs are close to one) but significant, which is a common situation in oncology.

18.4.3 Assessing Heterogeneity

As the different included studies are not conducted according to the same protocols, they will differ in at least a few aspects. Therefore, a certain level of heterogeneity

across studies is usually present, and it can be clinical, methodological, or statistical:

– *Clinical heterogeneity* is due to variability in the included population (e.g., participants' age, performance status, and prior treatments), variability in interventions (different drugs, different dose reduction management of the intervention), and variability in outcome (different definitions of an outcome).
– *Methodological heterogeneity* is due to variability in the risk of bias and/or variability in study design.
– *Statistical heterogeneity* is the variation in the treatment effects of the intervention being evaluated across the studies, i.e., the observed intervention effects are more different from each other than one would expect due to random error (chance) alone. Statistical heterogeneity arises as a consequence of clinical and/or methodological heterogeneity.

Graphically, statistical heterogeneity is presented as CIs from each study with poor overlap. There are statistical tests that can evaluate the heterogeneity between studies. The Chi-square (χ^2, Chi^2, or Q) is one of these tests and it measures how much the difference between effect measures is attributable to chance alone. However, this test has some expressive limitations, such as not being sufficiently powered to detect heterogeneity when few studies are included or when the studies have insufficient sample sizes. Also, as clinical and/or methodological variability often exists [36], some authors argue that detecting statistical heterogeneity could be pointless, since it will be present regardless of whether a statistical test is able or not able to detect it [37]. Therefore, quantifying the heterogeneity may be more useful than simply defining whether it is present or not. The Higgins (or I^2) inconsistency test describes the percentage of variability in the estimate of effect that is attributed to heterogeneity rather than chance. There are different recommendations on how interpret the result of an I^2 test. We suggest the following [37]:

- 0–25%—mild, acceptable heterogeneity
- 25–50%—moderate heterogeneity
- > 50%—high heterogeneity.

When heterogeneity is found, the authors have some options to deal with it:

– Use sensitivity analysis, subgroup analysis, or meta-regression.
– Do not perform a meta-analysis. The authors should only combine studies that are similar enough to be comparable. Although such decisions require qualitative judgments, when heterogeneity is significant and cannot be explained by any sensitivity analysis, the performance of a meta-analysis is not recommended.

18.4.4 Sensitivity and Subgroup Analyses and Meta-Regression

Sensitivity analysis involves repeating the meta-analysis after removing one or a few studies that met the included criteria. Any source of heterogeneity can be the subject of sensitivity analysis to explore its possible influence on the estimation of the effect. Also, sensitivity analysis can be done to find the source of statistical heterogeneity. It is also particularly useful for dealing with outliers, which often overestimate the effect of the intervention.

Subgroup analysis involves dividing studies, or the studies' participants, into subgroups according to clinical or methodological characteristics they share. Subgroup analysis of subsets of participants is almost always only possible in individual patient data meta-analysis (see Sect. 18.5). Although each subgroup can be more homogeneous than the entire group, the reader must be aware that subgroup analysis has limitations. First, it decreases the power of the analysis, since each subgroup has fewer studies and patients than the total of the subgroups, which can lead to a false-negative result in a subgroup. Second, the higher the number of subgroups analyzed, the greater will be the likelihood that one of them yields false-positive results. Finally, splitting patients from different studies into subgroups is not based on randomized comparisons, i.e., several other variables may be different and not balanced among patients in a subgroup and, hence, the findings may be misleading.

Meta-regression is a statistical test, similar to multiple regression, which aims to predict the effect estimate according to the characteristics of studies. The advantage of meta-regression over subgroup analysis is that the effect of multiple factors that might have modified the effect estimate can be analyzed simultaneously. However, the number of variables that can be considered to explain effect changes is limited by the number of studies available. Because of this, the Cochrane handbook recommends that "meta-regression should generally not be considered when there are fewer than ten studies in a meta-analysis." [24].

18.4.5 Understanding a Forest Plot

The most usual and informative way to present the results of a meta-analysis is in the form of a graph called a forest plot. This presentation shows the effect estimate and the CI for each study and for the meta-analysis, in addition to allowing rapid inspection of the studies' data and the conclusion of the meta-analysis. Different statistical software can yield forest plots with few differences. Also, the same software is capable of generating forest plots with different information, depending mainly on the type of data and the measure of effect used, as well as what the statistician wants to show. However, all forest plots share the same concepts of presentation.

For didactic purposes, we divided our forest plot [38] into three zones (Fig. 18.3a). In Fig. 18.3a, for zone 1, each line corresponds to a study, which is usually identified in the first column by author name and year of publication or

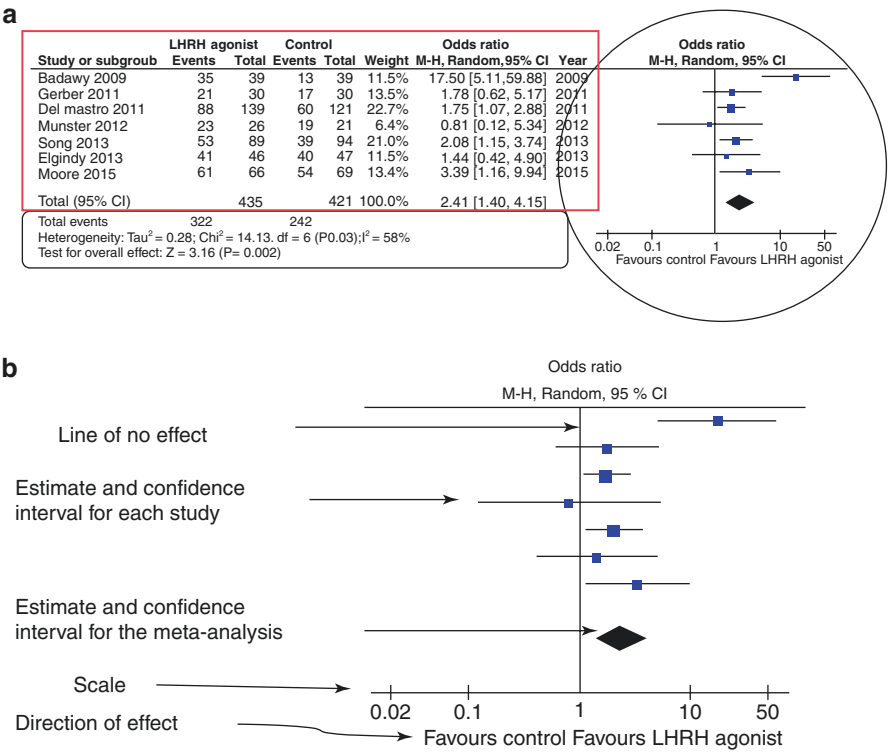

Fig. 18.3 An example of a forest plot, divided into three zones (**a**) for didactic purposes: the top-left zone (1—red rectangle) provides descriptive data from each study; the right zone (2—circle) presents the graphical nature of the information in zone 1, and the bottom zone (3—black rectangle) shows further statistical components of the forest plot. (**b**) Example of meta-analysis forest plot for interpretation. Courtesy of Munhoz et al. Gonadotropin-releasing hormone agonists for ovarian function preservation in premenopausal women undergoing chemotherapy for early-stage breast cancer: A systematic review and meta-analysis. JAMA Oncol. 2016;2 (1):65–73

the study's acronym. The information in the next columns may vary depending on the type of data. In our example we listed the event rates of each study (number of events in the total of patients in the intervention and control groups). If the meta-analysis had analyzed diagnostic tests, for instance, the forest plot could inform you of the true positives and true negatives instead. The second and third columns in Fig. 18.3a show the weight and the effect estimate of each study (the study's result), along with each study's corresponding 95% CI. The weight is related to the area of the squares in zone 2. The study estimate with it's corresponding 95% CI determines the position where the squares are and the width of the line on both sides. Usually, the bigger the square the smaller the lines. The meta-analysis (the overall effect estimate) is the black diamond that appears below the estimates of the included studies, where its edges correspond to its 95% CI. It is related to the last line of zone 1 (shown in bold).

For the interpretation of the graph in zone 2 (Fig. 18.3b), in addition to the information above, it is important to check the scale, where we will usually find the direction of the effect. Here, studies that concentrate the black squares to the left of the solid vertical line (the line of no effect) indicate results in favor of the intervention and the studies that concentrate their black squares to the right of the vertical line indicate results in favor of the control group. The same applies to the interpretation of the meta-analysis (diamonds). If the diamonds or lines representing the confidence intervals of each individual study are above the vertical line of absence of effect, the interpretation is that there are no statistically significant differences between treatments, or that the meta-analysis is inconclusive. Note that, when dealing with RRs, ORs, or HRs, the absence of effect is represented as 1. When dealing with mean difference, the absence of effect will be represented as 0 (as in Fig. 18.2c).

In zone 3, the first line simply summarizes the total of events. The last line gives you the *p*-value of the meta-analysis. Note that, as the diamonds do not cross the line of no effect, an overall effect *p*-value <0.05 is expected if the CI is defined as 95%. The second line shows data regarding heterogeneity analysis (see Sect. 18.4.3).

18.5 Individual Patient Data (IPD) Reviews

Rather than extracting data from study publications, the original research data may be available directly from the researchers responsible for each study. Individual patient data (IPD) reviews, in which data are provided on each of the participants in each of the trials, are considered the gold standard in terms of availability of data [39]. IPD minimizes the risk of bias and errors resulting from inadequate censoring. IPD can be re-analyzed centrally and eventually also combined in meta-analyses. On the other hand, IPD is usually more costly and time-consuming to obtain than other data. In addition, sometimes the data of all studies that meet the inclusion criteria cannot be available and analysis of only the available data entails a risk of selection bias.

IPD is particularly useful in oncology, where controversial questions and small benefits from interventions are common, and long-term follow-up for time-to-event endpoints (such as OS) is usually required. Situations where publications analyses are based on evaluable patients (not on all patients randomized), or situations where the published information is inadequate or where more complex statistical analysis is required are also well suited for IPD.

18.6 How to Present a Systematic Review with Meta-Analysis

After conducting all systematic steps, before submitting or presenting a review, it is important to return to the original question, and assess how well it was answered by the found evidence. Usually, it is important to evaluate how important the study design flaws are in the interpretation of the meta-analysis. When further research is

needed, some specific suggestions can be made about specific design features (better than a simple call for more data).

To assess the applicability of the results the authors should evaluate the inclusion/exclusion criteria. But it is also important to consider how a specific group would differ from the general population.

Presenting a systematic review with meta-analysis is more than just showing the numbers. We suggest a critical assessment, weighing up the beneficial and harmful effects of the interventions evaluated.

The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) Working Group presents a tool that helps to rate the certainty of the evidence found and the strength of final recommendations [40, 41]. GRADEpro, which can be found on the web (www.gradepro.org), is free and easy to use for summarizing and presenting information.

A systematic review should summarize the evidence in a clear and logical order. The authors can use a variety of Tables and Figures to present information, but we suggest following the PRISMA statement [21, 22] to improve the quality of reports.

Conclusions

As we have seen, through a rigorous methodological process, systematic reviews and meta-analysis help providers to keep up with the enormous amount of research data, judge the quality of studies, and integrate findings. Systematic reviews and meta-analysis yield greater precision of effect estimates, improve external validity (generalizability), providing consistency of results over different study populations, highlight the limitations of previous studies, and contribute to a higher quality of future studies. However, there are many points where authors should be careful in order to not add bias to their analysis and conclusions. Meta-analysis of randomized controlled trials with homogeneity is considered the highest level of evidence [42], but the situation where large randomized trials contradict a prior meta-analysis is still a field of debate [43–45].

References

1. Mulrow CD. Rationale for systematic reviews. BMJ. 1994;309(6954):597–9.
2. Tebala GD. What is the future of biomedical research? Med Hypotheses. 2015;85(4):488–90.
3. Pereira AA, Rego JF, Munhoz RR, Hoff PM, Sasse AD, Riechelmann RP. The impact of complete chemotherapy stop on the overall survival of patients with advanced colorectal cancer in first-line setting: a meta-analysis of randomized trials. Acta Oncol. 2015;54(10):1737–46.
4. Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R, et al. Should meta-analysts search EMBASE in addition to Medline? J Clin Epidemiol. 2003;56(10):943–55.
5. Sasse AD, Santos L. Searching LILACS database is irrelevant in systematic reviews in oncology. In: Evidence in the era of globalisation. Abstracts of the 16th Cochrane colloquium. Freiburg, Germany. p. 2008.
6. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA. 2004;291(20):2457–65.

7. Lee K, Bacchetti P, Sim I. Publication of clinical trials supporting successful new drug applications: a literature analysis. PLoS Med. 2008;5(9):e191.
8. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One. 2008;3(8):e3081.
9. Kicinski M. Publication bias in recent meta-analyses. PLoS One. 2013;8(11):e81823.
10. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629–34.
11. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50(4):1088–101.
12. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56(2):455–63.
13. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. J Clin Epidemiol. 2005;58(9):894–901.
14. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343:d4002.
15. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev. 2007;2. MR000010
16. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet. 2000;356(9237):1228–31.
17. Fergusson D, Laupacis A, Salmi LR, McAlister FA, Huet C. What should be included in meta-analyses? An exploration of methodological issues using the ISPOT meta-analyses. Int J Technol Assess Health Care. 2000;16(4):1109–19.
18. Burdett S, Stewart LA, Tierney JF. Publication bias and meta-analyses: a practical example. Int J Technol Assess Health Care. 2003;19(1):129–34.
19. Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, et al. Should unpublished data be included in meta-analyses? Current convictions and controversies. JAMA. 1993;269(21):2749–53.
20. Tetzlaff J, Moher D, Pham B, Altman D, editors. Survey of views on including grey literature in systematic reviews. 14th Cochrane Colloquium; Dublin, Ireland. 2006.
21. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.
22. Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62(10):1006–12.
23. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;7:10.
24. Higgins JPT, Green S (editors). Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The cochrane collaboration, 2011. Available from http://handbook.cochrane.org.
25. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol. 2006;59(7):697–703.
26. Keene ON. Alternatives to the hazard ratio in summarizing efficacy in time-to-event studies: an example from influenza trials. Stat Med. 2002;21(23):3687–700.
27. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials. 2007;8:16.
28. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet. 1998;352(9128):609–13.

29. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.
30. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials. 1996;17(1):1–12.
31. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol. 1998;51(12):1235–41.
32. Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, Gotzsche PC. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. Int J Epidemiol. 2007;36(4):847–57.
33. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;336(7644):601–5.
34. Holmes J, Herrmann D, Koller C, Khan S, Umberham B, Worley JA, et al. Heterogeneity of systematic reviews in oncology. Proc (Bayl Univ Med Cent). 2017;30(2):163–6.
35. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods. 2010;1(2):97–111.
36. Higgins J, Thompson S, Deeks J, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. J Health Serv Res Policy. 2002;7(1):51–61.
37. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557–60.
38. Munhoz RR, Pereira AA, Sasse AD, Hoff PM, Traina TA, Hudis CA, et al. Gonadotropin-releasing hormone agonists for ovarian function preservation in premenopausal women undergoing chemotherapy for early-stage breast cancer: a systematic review and meta-analysis. JAMA Oncol. 2016;2(1):65–73.
39. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. Eval Health Prof. 2002;25(1):76–97.
40. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924–6.
41. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol. 2017;87:4–13.
42. Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, et al. Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. Evidence-Based Medicine Working Group. JAMA. 2000;284(10):1290–6.
43. Villar J, Carroli G, Belizan JM. Predictive ability of meta-analyses of randomised controlled trials. Lancet. 1995;345(8952):772–6.
44. Cappelleri JC, Ioannidis JP, Schmid CH, de Ferranti SD, Aubert M, Chalmers TC, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? JAMA. 1996;276(16):1332–8.
45. LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med. 1997;337(8):536–42.