



# Reading and analyzing the medical literature should be methodic

Shuchun Li<sup>1</sup>, Min-Hua Zheng<sup>1</sup>, Abe Fingerhut<sup>1,2</sup>

<sup>1</sup>Department of General Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China; <sup>2</sup>Medical University Hospital Graz, Graz, Austria

**Contributions:** (I) Conception and design: A Fingerhut; (II) Administrative support: MH Zheng; (III) Provision of study materials or patients: A Fingerhut, S Li; (IV) Collection and assembly of data: None; (V) Data analysis and interpretation: None; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

**Correspondence to:** Abe Fingerhut, MD, FACS (hon), FRCS (g). Department of General Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China; Medical University Hospital Graz, Graz, Austria. Email: abefingerhut@aol.com.

**Abstract:** In the intricate landscape of diagnosing and treating diseases, healthcare professionals often rely on personal experiences and opinions. However, these sources may not always reflect the most current or rigorously tested approaches. Therefore, obtaining updated and evidence-based information from the current medical literature is essential for effective patient management. Yet, the vastness of the literature poses a challenge, as it encompasses a myriad of information that varies widely in quality and usefulness. Navigating this extensive sea of information requires an analytical and methodological reading approach. To discern the quality and utility of the information, readers should pose critical questions: (I) what are the results presented in the literature? (II) Are these results methodologically valid? (III) How can these findings be applied to a specific patient's case? (IV) How can you read and critically appraise a study according to its type? Understanding the type of study is crucial in evaluating the level of evidence. Studies can range from case series, cohort, and case-control studies to randomized controlled trials (RCTs), each offering a different level of reliability and applicability. Finally, to effectively sift through the literature and extract the most pertinent information for optimal patient care, a systematic approach to critical appraisal of medical literature is essential. This involves evaluating the study design, methodology, statistical analyses, and potential biases. By critically appraising the literature, healthcare professionals can ensure that the information they rely on is not only current but also methodologically sound, allowing for the best possible outcomes in patient treatment. The process of critical appraisal empowers clinicians to make informed decisions based on the highest quality evidence available, ultimately enhancing the quality of patient care.

**Keywords:** Medical literature; evidence-based medicine (EBM); critically appraise; P values; methodology

Received: 06 May 2023; Accepted: 21 December 2023; Published online: 15 January 2024.

doi: 10.21037/ales-23-29

**View this article at:** <https://dx.doi.org/10.21037/ales-23-29>

## Introduction

As physicians, we want to choose the best possible treatment for our patients, based on the correct diagnosis and the most appropriate and effective management plans. This decision should be based not only on the expected benefits but also taking into account potential adverse effects, patient status and fitness, and the setting in which the physician works. For the surgical community, health care professionals and the health care system in general: to obtain information that

is methodologically sound and critical is essential to provide the best possible care (and do no harm) for our patients.

Evidence-based medicine (EBM), a term coined by Guyatt and Sackett, is an approach that integrates the conscientious, explicit, and judicious use of current best evidence into everyday medical practice (1,2). It combines techniques from various disciplines such as science, engineering, biostatistics, and epidemiology, such as meta-analysis, decision analysis, risk-benefit analysis, and randomized controlled trials (RCTs), to ensure the delivery

of the right care to the right patient at the right time (3-5).

Physicians must persistently strive to access pertinent information from the continuously expanding body of literature in clinical research and practice to integrate the most reliable evidence into their daily practice (6). Physicians in different parts of the world approach and read medical literature in various ways. Doctors in Western countries, such as the USA and Europe, prioritize EBM and clinical research, relying on resources like PubMed for the latest medical literature. In Asian countries like China and India, doctors may incorporate traditional medicine alongside modern practices. While some still hold traditional beliefs, an increasing number focus on EBM and clinical research. Reading literature is a crucial means to stay updated on the latest medical advancements, and this practice is consistent worldwide (7,8). Therefore, scrutinizing the literature to find the information with the highest level of evidence is an important skill that is learned from methodic reading and analysis of the medical literature, a skill also called critical appraisal (9,10).

To determine which diagnostic method is the most efficient and best fits the situation, which is the most appropriate for a specific patient's disease, or what to make of comparisons between two or more different diagnostic or therapeutic methods, an analytic and methodologic approach is needed. We must address several questions: (I) what are the results in the paper under analysis? (II) Are they methodologically valid? (III) How can these results be of use to a particular patient? In the current manuscript, we would like to answers to these three questions which may lead us to understand the results, critically appraise the methods that led to the results, and then, evaluate how the authors validated and compared their result to those in the literature, and finally, based on the data and appraisal, learn how to best adapt the correct line of thought for the particular problem.

## Review

### *What should be emphasized in the results section of the research paper?*

Aside from descriptive statistics, this requires determination of the size and precision of the treatment effect. The effect size must be analyzed within the context of the problem we are faced with: this context can be found in the introduction of the paper. Effectively, the introduction is where the reader discovers the background of the problem, what has

been done or what is the current thought with respect to the problem. Effect size is the quantitative measure of the magnitude of the experimental effect. The larger the effect size the stronger the relationship between the variables or parameters under investigation. Effect sizes either measure the degree of associations between variables or the proportions of differences between group means. This will be found in the methods and results chapters and often in Tables and Figures that provide a visual aid, complementary to the text.

### *Determine the rationality of the approach used in a research paper*

The validity of the results means checking that the results of the study correspond to the direction (better or worse) and the magnitude of the underlying true effect that is observed in the study. The methodology used to obtain the results has to be scrutinized. What is the population in the study? Are patients included without selection bias? What statistical test was used to validate the credibility of the results? Is the type of study and the methodology adapted to the problem? Have confounding factors been dealt with and how? Have missing values been accounted for and how? Is the diagnostic test well described, in a reproducible fashion?

Low P values (said to be statistically significant when they are below the legendary but widely admitted cut-off of less than 0.05) are commonly interpreted as a proof of a strong relationship between two variables. This is ignoring that P values are probabilities, the result of a statistical test, and not a label of truth or proof. They just give the probability of obtaining at least as extreme a result if the experiment were conducted again, provided the likelihood that null hypothesis is true (11). While a P value can indicate the probability of an intervention being effective, effect sizes are needed to tell us how effective it is. Of note, effect size is independent of sample size, to the contrary of P values.

### *Assessing the applicability of research results*

Can these results be applied to patients under the care of the reader is equivalent to first determining the external validity, i.e., whether the results of the study are generalizable, and secondly, whether they are credible and applicable to the population to be treated. How do the results compare to what has already been done? Are the populations in the literature the same as in the study? Does

the literature reinforce or counter the findings of the study under analysis? How can the differences be explained? With this information, can the results be applied to the readers' population? This can be found in the discussion. If not, it is up to the reader to analyze and critically appraise the literature to find the answers.

The reader should be able to easily determine the type of study without the need to search for this information within the text. Typically, the type of study is indicated in the title or can be identified through the keywords used. This ensures that readers can quickly understand the study design and methodology. Studies can be observational (that draw inferences from a sample to a population where the independent variable is not under the control of the investigator) or experimental (that draw inferences from a sample where the independent variable is under the control of the investigator). The literature commonly includes four types of studies, including case series (non-comparable studies, akin to audits), cohort and case control studies (comparative observational studies) and controlled (clinical) randomized trials (comparative experimental studies) (12). Randomized trials are a form of scientific experiment used to account for factors that are not under direct experimental control (the control group is used for comparison because the participants are not exposed to and/or do not experience the effect of the independent variable. Among non-randomized observational studies, when patients are followed forward in time, starting from the exposure and tracking them until the consequences of the exposure (target outcome) occur, this is called a cohort study (they are usually prospective in that all the subjects are outcome-free at the start; some also define retrospective cohort studies, where at least some of the subjects have already developed the outcomes). When the investigator has determined a group with a specific target outcome and another without this same outcome, and then goes back in time to study exposures or exposure factors, case control study. Examples of randomized controlled studies include clinical trials that compare the effects of drugs, medical devices, diagnostic procedures, or surgical techniques or management strategies where the participants are allotted to one or the other treatment arms (one with the usually new treatment to assess, called the experimental arm, the other with a standard, called the control arm) by a randomization process, currently, usually computer generated. They can be used to evaluate treatment effects, typically those considered to be beneficial. However, when researchers aim to assess harm, randomization is often considered unethical.

In such cases, non-randomized observational studies can be utilized by examining whether patients have been exposed to a harmful agent, either through predetermined factors or by chance.

### *Read and critically appraise a study according to its type*

#### **Case series**

Case series are observational studies that provide data from a selected group of participants without a control (comparison) population (13). Case series are highly prone to bias because they represent the works of a predefined (usually single) team, with their proper indications, techniques and evaluation procedures (14). The level of evidence is low (level IV) (15). Nonetheless, the literature abounds and most often, when the authors are well known, the results are regarded as being on a much higher level. Information reported in case series can be used as audits, or sometimes to describe outcomes of novel treatments; in this case, they can be used to generate hypotheses to describe new techniques or treatment protocols before future studies with stronger trial design (14). Their main advantage is that they are easy to conduct, require less time and financial resources than RCTs, case-control, or cohort studies. Chan (14) summarized the key points of a good case series as follows: (I) clear study objective/question; (II) well-defined study protocol; (III) explicit inclusion and exclusion criteria for study participants; (IV) specified time interval for patient recruitment; (V) consecutive patient enrollment; (VI) clinically relevant outcomes; (VII) prospective outcome data collection; (VIII) high follow-up rates. Guidelines have been published (16).

For the latter three, guidelines have been published for different types of studies on the EQUATOR (Enhancing the QUALity and Transparency Of health Research) web site (<https://www.equator-network.org/>). The site exists in several languages, including Chinese (<https://www.equator-network.org/about-us/chinese-equator-centre/>). One of the initial steps of analysis is to make sure the article under consideration follows these guidelines.

#### **RCTs**

In a RCT, the randomization process should determine whether the patients included have similar base-line characteristics, similar prognostic or risk factors, and that the only difference between the two populations concerns the treatment (to be tested) (17). The reader should control whether the three following clauses have been

respected: uncertainty, equipoise and ambivalence. The uncertainty principle states that physicians should enroll a patient only if they are substantially uncertain which of the treatments is most appropriate (18); equipoise means there is absence of consensus within the expert clinical community about the comparative merits of the alternatives to be tested (19) and finally, ambivalence means that each and every patient included must be able of receiving any one of the treatments independently (20). The last clause means that randomization has to take place at the latest moment possible, when the investigator has eliminated all exclusion criteria and all other criteria have been met (20). If the number of patients in a study is sufficiently large, it is expected that most, if not all, confounding factors (both known or unknown) would be distributed evenly between the control and experimental treatment arms. This balance in confounding factors helps minimize their influence on the study outcomes. Other methods that can be used to balance the risk factors before randomization include stratification and randomization schemes by blocks.

#### *Were the patients truly randomized?*

There is a difference between RCT where the sequence generation is achieved using computer random number generation or a random number table and a controlled clinical trial (CCT) where allocation not strictly speaking random (e.g., patients are allocated to one arm or the other by tossing a coin, using dates, names, or admittance numbers or order to determine which exposure will be allotted to which group). True randomization refers to the process of allocating participants to different treatment groups in a study without any bias. It ensures that the allotment sequence, such as the choice of administering one treatment or another to be compared, is free from systematic influences or favoritism (20). The reader should carefully examine the methods section of a study to ensure that the authors have provided sufficient information regarding the randomization process. Specifically, the reader should look for the following characteristics: (I) consecutive sequence (all eligible patients were included in the randomization process), (II) unpredictable sequence generation (the randomization sequence was generated using a method that ensures the next element of the sequence is impossible to know in advance. This implies that methods of allotment based on factors such as date of birth, date of entry to the hospital, or order of entry to the study are not considered valid, as these can introduce bias), (III) allocation concealment (the actual treatment assignment for a particular patient remained unknown to both the

patient and the person administering the care. This ensures allocation concealment and prevents any potential biases or influences related to treatment knowledge), (IV) blinded outcome assessment [the reader should assess whether the outcome was evaluated by someone who was unaware of the treatment arm to which each patient was allocated (21)]. Finally, the reader should be sure that blinding was performed in such a way as to ensure that the patients are evaluated in the same manner (no bias of selection, follow-up, classification nor evaluation).

Ideally, when both the patients, the assessor, and the care provider (surgeon) are unaware of the assigned treatment, it is referred to as a “triple-blind” study. Obviously in surgery, the surgeon cannot ignore the treatment. However, double or single blinded studies should not be rejected or thought to be inferior. Greater credibility should be placed in results when at the least the assessor of outcomes was blinded to the method used. To ensure the proper implementation of blinding in a study, explicit information should be provided in the methods section regarding who was blinded and how the blinding was conducted. Relying solely on the term “blinded” without further clarification may not provide enough evidence of effective blinding and bias reduction.

Lastly, the reader should assess whether the follow-up in the study was complete. If there were any instances of incomplete follow-up, it is important to consider the potential impact of this incompleteness, known as attrition bias, on the study outcomes (22). For example, during the follow-up of a study on a disease with recurrence as the main outcome measure, patients who are lost to follow-up may have experienced various circumstances. These can include cases where patients have passed away, relocated to a different area, or did not respond to recall invitations. However, if such information is not provided in the study report, it is possible that patients who were dissatisfied with the outcome sought medical advice from another surgeon. To ensure the validity of study results, an informal rule recommends that the number of unaccounted for patients at the time of assessment should not exceed 10%. This guideline serves as a benchmark for evaluating the completeness of follow-up in a study. If the percentage of unaccounted patients exceeds this threshold, alternative methodologies such as the “maximal bias” or “worst-case scenario” approaches can be employed (23). This entails categorizing the outcomes of all patients in the group with more favorable results, but whose long-term status is uncertain, as either a poor result or a failure. If analyzing the data in this way does not impact the study findings, it can be

inferred that patients lost to follow-up did not significantly affect the results (bias). However, if the outcome is altered, it impedes the ability to draw valid conclusions. It's crucial to acknowledge that an increased number of patients lost to follow-up diminishes the study's validity. Regrettably, such analyses are infrequently performed. Readers must be mindful that without such scrutiny, the reliability of the study's conclusions should be approached cautiously.

#### **Critical importance of evaluating and interpreting research conclusions**

One particularly salient aspect of critical appraisal is to evaluate the contents and value of the conclusion. The conclusion should reflect the impact of the results on the overall population and how they can be employed in other settings. If an effect causal can be applied (correctly performed randomized trial with a little bias as possible), then a declarative language can be used to express the results (24). However, if the results are not conclusive (no statistically significant difference) or underpowered, only descriptive terms should be used (24). Moreover, the reader should be able to detect what Boutron *et al.* have called "SPIN" (25,26). This has been defined as a reporting (writing) strategy that aims to say that the experimental treatment is beneficial although no statistically significant difference was found for the primary outcome, or even to mislead (distract) the reader from statistically non-significant results. This also includes subgroup analysis, coming to a conclusion on a particular (sub) group of the studied population that was not pre-defined in an unbiased fashion (RCTs) and for which the study is almost always underpowered. This is why we believe reading the conclusion before analyzing the article, or only the abstract, can be misleading.

#### **Critical appraisal of results**

Analysis of the results of a study should include assessing both the magnitude of the treatment effect and the precision of the findings. One approach to determining the magnitude of the treatment effect is by calculating the absolute risk reduction (ARR) or risk difference (27). As an example, let's consider a scenario where the recurrence rate with treatment A (control group) is 15% (x), while the recurrence rate with treatment B (treatment group) is 10% (y). The ARR is calculated as  $x - y = 0.15 - 0.10 = 0.05$ . Alternatively, we can quantify the treatment effect's magnitude using the relative risk (RR), comparing the recurrence risk between patients receiving treatment A and those receiving treatment B. The RR is calculated by dividing the risk in the treatment group (y) by the risk in

the control group (x). The RR would be  $y/x = 0.10/0.15 = 0.67$ . When expressing results, the reader should ensure that numbers are provided: simple percentages may be misleading.

The relative risk reduction (RRR) is indeed a commonly used measure for dichotomous outcomes, such as treatment effect (yes or no) or survival (dead or alive). It represents the complement of the RR and is expressed as a percentage. This is expressed as a percentage  $(1 - y/x) \times 100\%$ , or in this case,  $(1 - 0.67) \times 100\% = 33\%$ .

One spin-off of the ARR is the number needed to treat to observe one positive (or adverse/harmful) event, also called clinical significance (28). It is simply calculated as  $1/ARR$ . In the above example, this would be  $1/0.05 = 20$ . In other words, one would have to treat 20 patients to observe one beneficial (or adverse) event (also called number needed to harm).

Determining the precise nature of the treatment effect, also known as the true risk reduction, can be challenging. The best estimation available is the observed treatment effect, often referred to as the point estimate. However, it is important to remember that this estimate is inherently imprecise. To capture this imprecision, researchers calculate confidence intervals (CIs), which provide a range of values within which the true population parameter is likely to exist. Typically, the 95% CI is reported, which signifies a range of values that includes the true risk reduction 95% of the time (29). In other words, if the study were repeated multiple times, we would expect 95% of those intervals to contain the true value in the population.

CIs are recognized for their quantitative value, in contrast to "P" values that offer a qualitative measure of probability rather than indicating the strength of evidence against the null hypothesis of "no effect" (11). The P value alone does not provide information about the magnitude or direction of a difference. In contrast, CIs offer valuable insights into the level of evidence regarding specific quantities of interest, such as the benefit of treatment (29). CIs hold significance in research findings and should be included in the main text and abstract of published articles that report results from RCTs and other studies.

CIs can also provide insights into the clinical significance of research findings (30). This larger number of events allows for a more precise estimation of the treatment effect, resulting in a narrower CI. A narrower CI indicates that we have greater confidence that the true RRR or any other measure of efficacy is closer to the observed value in the study. CIs can also indicate the direction of the effect.

Indeed, if a CI has a lower limit that falls below zero, it suggests that the treatment effect could potentially be harmful. The reader must check that the sample size (power calculation) was performed correctly (delta or difference that the analysis of the literature seems to indicate as being adequate, or clinically appropriate along with the risks “alpha”, “beta”, or also called type 1 or type 2 errors. The required sample size for a RCT can be determined using various formulas, which are readily available in most computer programs. It is essential for researchers to calculate and report the calculated sample size in the methods section of their paper.

When considering clinical significance, particularly in a positive study, it is important to examine the lower limit of the CI. If the lower limit of the CI still aligns with a RRR that is considered clinically meaningful and effective for recommending the treatment to patients, it suggests that the number of patients enrolled in the study was adequate. However, if the lower limit of the CI falls below a threshold considered clinically relevant, it implies that the treatment effect may not be substantial enough to warrant recommendation, despite statistical significance (11,31). In a negative study, conversely, examining the upper limit of the CI helps determine its clinical relevance. If the upper limit is clinically relevant, it implies that not only did the study fail to demonstrate the superiority of the experimental treatment over the control modality, but it also failed to provide evidence that the experimental treatment is not better.

Finally, the reader needs to verify if the patients were analyzed based on their initially assigned groups, regardless of whether they actually received the assigned treatment. This principle is known as the “intention-to-treat” principle. There are two key reasons for the significance of this analysis. Firstly, if a patient switches from one arm to another due to unexpected challenges or pathological results, analyzing outcomes based on the actual treatment received could introduce bias in favor of the group without difficulties or with “normal” pathological findings. Secondly, unexpected difficulties or pathological findings are common occurrences in clinical practice and should be considered as part of the overall treatment process.

#### ***Methodology considerations***

The statistical tests employed to compare differences between two groups, whether it pertains to demographic data or results, depend on the nature of the data being analyzed. Specifically, it hinges on whether the data is continuous, represented by numerical values on a scale, or

categorical, characterized by binary outcomes such as yes/no or alive/dead (31). One of the most commonly observed errors is to see durations (operation, hospital stay, period of recuperation...) expressed as means with standard deviations; these are often not normally distributed data, especially when the overall population is small. This error is not misleading in itself, but when a parametrical statistical test is applied (where a non-parametrical statistical test is called for), the result no longer has any meaning and can be misrepresented.

When assessing normally distributed “continuous” parameters, the Student *t*-test is the most commonly used statistical test. However, if the data distribution is not normal, an alternative test known as the Mann-Whitney *U* test should be employed (32).

Where values are derived from the same patient (paired data), if for instance, one intends to assess the tumor diameter in for individuals diagnosed with rectal cancer undergoing neoadjuvant radiotherapy, the *t*-test can be employed when there are at least 20 measurements both before and after the treatment, and the distribution of the data is normal. However, if the data distribution is not normal or Gaussian, the more appropriate approach would be to employ the Wilcoxon matched-pairs test.

When dealing with “categorical” outcomes, the commonly employed statistical test is Pearson’s chi-squared test. However, if the number of measurements is less than 20, Fisher’s exact test is more appropriate. In situations where the data is paired, such as in before-and-after comparisons, McNemar’s test should be utilized.

When comparing the average number in three groups of patients undergoing different treatment(s), the choice of statistical test depends on the nature of the data and the study design. The one-way analysis of variance (ANOVA) should be used for comparing means among the three groups. On the other hand, if the data does not follow a normal distribution or the assumptions for parametric analysis are not met, the Kruskal-Wallis test can be used.

#### **Other (non-randomized) observational studies**

In case control and cohort studies, readers should consider the following three questions: “What are the (exact) results?”, “Are the results valid?”, and “How can I apply them to my patient(s)?”

#### ***Cohort studies***

When prospective, this type of study consists of identifying two groups, one exposed and the other nonexposed. The analysis involves following participants forward in

time and monitoring their outcomes. Cohort studies are particularly useful for assessing rare events, such as potential harm. Conducting a RCT for such events can be ethically questionable or even impossible due to the need for a large number of participants, and informing subjects about the potential harm. The risk of bias in these studies is considerable. It is crucial for the reader to evaluate whether the two groups (exposed and non-exposed) started the study with disparate risks for the outcome or if other associated factors (potential confounding variables) influenced the choice of treatment. It is essential to verify if these differences were documented and analyzed, allowing the assessment of how dissimilar the two groups were regarding all factors except the exposure. Additionally, the reader should confirm that the authors employed appropriate statistical techniques to account for these differences, if applicable. In contrast to RCTs, where little-known confounding factors are expected to be evenly distributed by chance after randomization, cohort studies may still be vulnerable to confounding bias, which might remain unnoticed. Consequently, the strength of inference in cohort studies is generally regarded as lower than that of a RCT.

In a retrospective cohort study of factors predicting the success of bariatric surgery as an example, one cohort study examined the excess weight loss after sleeve gastrectomy (SG) or Roux-en-Y gastric bypass (33). The study found a lower baseline body mass index (BMI) and absence of type 2 diabetes (T2D) were predictive of therapeutic success and %EWL. Since patients who underwent SG had a higher baseline BMI and lower T2D rate, it would be incorrect to conclude that the success rate after surgery was related to the surgical method.

#### **Case-control studies**

When the event of interest is rare or takes a long time to occur, a case-control study is an alternative investigative technique. In this study type, patients who have encountered the target outcome are compared to a control group that shares similar demographics, like age, sex, and prognostic factors, but has not experienced the target outcome. The reader should look for an assessment of the relative exposure frequency in both groups. It's crucial to check if the authors have addressed differences in known and measured prognostic factors between the cases and controls. A thorough analysis should encompass all potential confounding factors that might affect the association.

As with RCT, if follow-up is not complete, missing patients need to be taken into account. Tallying the missing outcomes is problematic as the methods to compensate this

lack of data are different from those used for RCT.

#### **Outcome measures**

The reader should evaluate the strength of association between the exposure and outcome being investigated. This is not the role of P values. The methods differ slightly between cohort and case-control studies.

In cohort studies, the RR is calculated to measure the association between exposure and outcome. If the RR is greater than 1, it indicates an increased risk of developing the outcome among the exposed individuals. Conversely, if the RR is less than 1, it suggests a decreased risk of developing the outcome among the exposed individuals. In case-control studies, direct calculation of RR is not possible since the investigator selects cases and controls, and the proportion of individuals with the outcome is not determined by chance. Instead, another measure called the odds ratio (OR) is commonly used to assess the association between exposure and outcome, the odds of having an event, divided by the odds of not having the event in patients being exposed or the probability of success divided by the probability of failure in patients with the same exposure is the effect measure of interest. This is typically done for retrospective studies.

The precision of the risk estimate is reflected by the width of the CI around the estimate. In a study that demonstrates an association, the lower limit of the CI for the RR estimate represents the minimum strength of the association. It provides the lowest possible estimate of the risk associated with the exposure. Conversely, in a negative study where no statistically significant association is found, the upper limit of the CI indicates the potential magnitude of the risk despite the lack of significant results.

#### **Meta-analysis**

In the last several decades, reviews of the literature have become omnipresent. They can be narrative, scoping or systematic, and should be announced as such in the title, and detailed in the methods according to the appropriate guidelines (<https://www.equator-network.org/>). Meta-analysis is a statistical procedure that amalgamates and summarizes outcomes for a particular variable obtained from multiple studies, typically akin in nature (34). These studies should be sourced from a systematic literature review, which ought to be delineated in a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram within the report detailing the meta-analysis (35,36). Furthermore, the protocol for conducting the meta-analysis should be registered with PROSPERO, an international registry dedicated to systematic reviews

in health and social care (35). While the goals of meta-analyses are well defined (increase the power (the chance of detecting a real (statistically significant) effect, if one exists, by combining individual studies that are too small to detect lesser effects or reach “statistical significance” or answer questions not posed by the individual primary), many of these publications today are methodologically flawed or misused (36,37). Differences in results have to be assessed and statistical analysis of findings allows the evaluation and quantification of the degree of disparity. Let it be understood that the use of statistical methods does not guarantee that the results of a review are valid, any more than it holds true for a primary study. Moreover, like any tool, statistical methods can be misused. For example, Totaro *et al.* conducted a systematic review of special complications related to minimally invasive gastric surgery approach and their impact on survival, they used a checklist in their report to make the review more scientific (38). This should be the rule, and the reader should look for this information when appraising the article.

The label “meta-analysis” should not be considered as being more reliable or automatically at the summit of the pyramid of evidence (39,40). Critical appraisal of meta-analyses requires looking at how the authors searched for and analysed heterogeneity (use of the Q, I<sup>2</sup> and Tau<sup>2</sup> statistics and eventually the use of prediction intervals), with proper explanations (41,42). Explanations of these statistics and the accepted values for pertinence can be found (37,42). Once again, we have to be able to distinguish justified conclusions from “SPIN” (25,26).

## Conclusions

Ultimately, a meticulous, systematic, and stepwise assessment, commonly known as “critical appraisal”, of all scientific papers is imperative to establish the validity (10), credibility, and generalizability of the information. This process is essential before readers can draw conclusions or infer any associative properties. Evidence may comprise empirical observations regarding the apparent relationship between events, and we have developed a hierarchy to categorize this evidence into different “levels of evidence”. However, it is not just the label of the level of evidence that counts. The attentive reader has to critically appraise to make sure that the methodology used by the author and the reported results are appropriate to the type of study and that the conclusions are drawn with reference to the methodology. This is the goal of methodic critical appraisal.

## Acknowledgments

*Funding:* This work was supported by National Natural Science Foundation of China (82072614 to M.H.Z.); Shanghai Key Clinical Discipline Construction Project (shslczdk00102 to M.H.Z.); GuangCi Deep Mind Project of Ruijin Hospital Shanghai Jiao Tong University School of Medicine (A.F.).

## Footnote

*Peer Review File:* Available at <https://ales.amegroups.com/article/view/10.21037/ales-23-29/prf>

*Conflicts of Interest:* All authors have completed the ICMJE uniform disclosure form (available at <https://ales.amegroups.com/article/view/10.21037/ales-23-29/coif>). M.H.Z. and A.F. serve as the co-Editors-in-Chief of *Annals of Laparoscopic and Endoscopic Surgery*. M.H.Z. reports that he received support from National Natural Science Foundation of China (82072614) and Shanghai Key Clinical Discipline Construction Project (shslczdk00102) for this work. A.F. reports that he received support from GuangCi Deep Mind Project of Ruijin Hospital Shanghai Jiao Tong University School of Medicine for this work. The other author has no conflicts of interest to declare.

*Ethical Statement:* The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

*Open Access Statement:* This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

## References

1. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA* 1992;268:2420-5.
2. Thoma A, Eaves FF 3rd. A brief history of evidence-based medicine (EBM) and the contributions of Dr David

- Sackett. *Aesthet Surg J* 2015;35:NP261-3.
3. Evidence-Based Medicine. Division of General Internal Medicine, Johns Hopkins Hospital. Available online: <https://www.hopkinsmedicine.org/gim/research/method/ebm.html>
  4. Dziri C, Fingerhut A. What should surgeons know about evidence-based surgery. *World J Surg* 2005;29:545-6.
  5. Guyatt G, Rennie D, Meade MO, et al. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. New York, NY: McGraw-Hill Education; 2015.
  6. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *Aust Vet J* 1996;74:311.
  7. Raza MA, Nor FM, Mehmood R. Reading habits of medical practitioners: Young doctors in Pakistan, a case study. *J Taibah Univ Med Sci* 2022;17:844-52.
  8. Burke DT, DeVito MC, Schneider JC, et al. Reading habits of physical medicine and rehabilitation resident physicians. *Am J Phys Med Rehabil* 2004;83:551-9.
  9. Burls A. *What is critical appraisal?* London: Hayward Medical Communications; 2009.
  10. Fingerhut A, Lacaine F. *Critical Appraisal*. *Surg Innov* 2017;24:101-2.
  11. Fingerhut A. Probability, P values, and statistical significance: instructions for use by surgeons. *Br J Surg* 2023;110:399-400.
  12. Horton R. The rhetoric of research. *BMJ* 1995;310:985-7.
  13. Mathes T, Pieper D. Clarifying the distinction between case series and cohort studies in systematic reviews of comparative studies: potential impact on body of evidence and workload. *BMC Med Res Methodol* 2017;17:107.
  14. Chan K, Bhandari M. Three-minute critical appraisal of a case series article. *Indian J Orthop* 2011;45:103-4.
  15. *Levels of Evidence (March 2009)*. Oxford Centre for Evidence-Based Medicine. 2009. Available online: <https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009>
  16. Agha RA, Sohrabi C, Mathew G, et al. The PROCESS 2020 Guideline: Updating Consensus Preferred Reporting Of CasE Series in Surgery (PROCESS) Guidelines. *Int J Surg* 2020;84:231-5.
  17. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. *BJOG* 2018;125:1716.
  18. Peto R, Baigent C. Trials: the next 50 years. Large scale randomised evidence of moderate benefits. *BMJ* 1998;317:1170-1.
  19. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987;317:141-5.
  20. van Werkhoven E, Tajik P, Bossuyt PM. Always randomize as late as possible. *Gastric Cancer* 2019;22:1308-9.
  21. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. *Lancet* 2002;359:696-700.
  22. Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomised controlled trials. *BMJ* 2006;332:969-71.
  23. Gamble C, Hollis S. Uncertainty method improved on best-worst case analysis in a binary meta-analysis. *J Clin Epidemiol* 2005;58:579-88.
  24. Fingerhut A, Sarr MG. Strength of verbs in medical writing should correspond to the level of evidence (or degree of causality): A plea for accuracy. *Surgery* 2017;161:1453-4.
  25. Boutron I, Dutton S, Ravaud P, et al. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 2010;303:2058-64.
  26. Fingerhut A, Lacaine F, Cuschieri A. Medical SPIN: misinformation by another name. *Surg Endosc* 2015;29:1257-8.
  27. Robbins AS, Chao SY, Fonseca VP. What's the relative risk? A method to directly estimate risk ratios in cohort studies of common outcomes. *Ann Epidemiol* 2002;12:452-4.
  28. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988;318:1728-33.
  29. Altman DG. Why we need confidence intervals. *World J Surg* 2005;29:554-6.
  30. Fingerhut A. Response to Dr Slim. *Surgery* 2023;173:1102-3.
  31. Greenhalgh T. How to read a paper. Statistics for the non-statistician. II: "Significant" relations and their pitfalls. *BMJ* 1997;315:422-5.
  32. Guller U, DeLong ER. Interpreting statistics in medical literature: a vade mecum for surgeons. *J Am Coll Surg* 2004;198:441-58.
  33. Grandt J, Chang J, Türler A, et al. Cohort study evaluating predictors of therapeutic success after sleeve gastrectomy or Roux-en-Y gastric bypass. *Ann Laparosc Endosc Surg* 2022;7:1.
  34. Paul J, Barari M. Meta-analysis and traditional systematic literature reviews—What, why, when, where, and how? *Psychology & Marketing* 2022;39:1099-115.
  35. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis

- protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2015;350:g7647.
36. Greenhalgh T. Papers that summarise other papers (systematic reviews and meta-analyses). *BMJ* 1997;315:672-5.
37. Andrade C. Understanding the Basics of Meta-Analysis and How to Read a Forest Plot: As Simple as It Gets. *J Clin Psychiatry* 2020;81:20f13698.
38. Totaro L, Celotti A, Ranieri V, et al. Specific complications related to the approach in minimasive gastric surgery and impact on survival: a narrative review. *Ann Laparosc Endosc Surg* 2022;7:22.
39. Cook DJ, Guyatt GH, Laupacis A, et al. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1992;102:305S-11S.
40. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1989;95:2S-4S.
41. Green S, Higgins JP, Alderson P, et al. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane Handbook for Systematic Reviews of Interventions. Chichester, UK: John Wiley & Sons; 2008.
42. Borenstein M. *Common Mistakes in Meta-Analysis: And How to Avoid Them*. New Jersey, USA: BioSTAT, Inc.; 2019.

doi: 10.21037/ales-23-29

**Cite this article as:** Li S, Zheng MH, Fingerhut A. Reading and analyzing the medical literature should be methodic. *Ann Laparosc Endosc Surg* 2024;9:8.