Critical Appraisal of the Medical Research Literature

K M Ariff, M.Med., General Practitioner, 02600, Arau, Perlis

Summary

The medical practitioner of today is often faced with the dilemma of having to cope with a constant stream of medical literature that arrives at the practice. Trying to keep abreast with recent advances in medical knowledge while maintaining a busy work schedule is a common problem for practicing health professionals. While review articles are usually read with relative ease, original research reports require some critical reading skills, especially if one intends to utilise the research findings in critical decision making. This article provides a systematic approach to critical reading and appraisal of a research article using the IMRAD format.

Key Words: Critical appraisal, Medical research.

Introduction

A commonly encountered dilemma in medical practice is in having to cope with the endless volumes of medical literature that arrives at the practice. This is especially seen in general practice where the practitioner is expected to keep abreast with advances in a diversity of disciplines while practice overload may limit his/her time to a valuable commodity. While review articles are often read with minimum effort, reading original research articles can make considerable demands on one's time and concentration. The reader can best identify the sort of articles that are worth reading, selecting useful ones while rejecting unhelpful ones without feeling guilty. If the purpose was only to obtain a broad overview of recent medical advances, then skimming the article without scrutinising the study design and methodology may be acceptable. But, if the reader was reviewing the literature in preparation for a research project or evaluating the results of a clinical trial to determine its validity in relation to patient care, then considerable caution is required. In such instances, reading the abstract and conclusion sections of the paper while ignoring the methods and results sections can sometimes result in poorly designed reports with inherent flaws and weaknesses deceiving the unwary reader and influencing his clinical decision making. It is worthwhile remembering that even the most reputed peer-reviewed journals have been known to publish controversial materials. This article is expected to provide some guidelines on the systematic approach to critical appraisal of research reports which, with some practice, can be an enormous asset to anyone with an interest in the subject.

Objectives of Critical Appraisal

Broadly, this involves a systematic and logical evaluation of a scientific article to determine its validity and relevance to clinical practice. Basically, the exercise focuses on three main issues. Firstly, the reader verifies whether the basic research information is provided in the paper. Secondly, one determines whether the manner in which the information was obtained is appropriate and accurate. Thirdly, the critical reader looks for other factors that may threaten the validity of the study.

The Structure of a Research Report

The IMRAD format has now become the accepted

format for writing scientific reports. IMRAD stands for the different sections of the article (Introduction, Methods, Results and Discussion). The introduction section is usually preceded by an abstract or summary section.

The Abstract

The abstract or summary should provide sufficient information so as to enable the reader to decide whether he/she has an interest in the subject of study. The abstract should state clearly the purpose of the study or investigation, basic procedures undertaken (selection of study subjects, study design and analysis), the main findings and the principal conclusions derived.1 If the subject of the study arouses the reader's interest, then he/she can best decide whether the entire article is worth reading. At this juncture, the reader can ask whether the results of the study are important and relevant to his/her practice.

The Introduction to the Report

This section should clarify the purpose of the article and summarise the rationale for the observation or study. The author should provide pertinent references that support the need for the study while providing relevant background information on the subject. The reader should ensure that the research questions, hypothesis (if any) and the objectives of the study are specified and consider whether the research question is answerable and is of sufficient practical importance. The number of research questions to be answered requires some scrutiny. This is especially pertinent in clinical trials that attempt to answer too many questions and often yield unsatisfactory results. The primary objective of the trial should be clearly specified while all other questions are considered to be of secondary importance. It is important for the critical reader to verify if the study is an original one or a repetition of a previous study on the subject. If a similar study has been conducted previously, the author should state clearly how his/her study differs from previous studies in relation to size, duration of study, methodology and the characteristics of the population studied.

The Methods Section

This section is an important component of a scientific journal article and shall be discussed in some detail here. Broadly, it provides information on how the study was conducted, the type of study design used and allows the reader to judge the validity and its findings.

The Study Design

Study designs can be broadly divided into experimental studies or clinical trials and observational studies which include descriptive studies, cohort studies, case-control, cross-sectional and case-series studies. Well designed clinical trials (experiments) have traditionally been regarded as providing the strongest evidence for causation with the least number of problems or biases.[2] Randomized controlled clinical trials have the strongest study design and ideally clinical trials should be double blind and randomized. Randomization safeguards against selection bias while the blinding process minimises or eliminates bias in the assessment of responses to treatment and other interventions. The possible sources of bias in clinical research are described below. Case-control studies investigate associations between exposures and a condition of interest. A study group with a certain condition (cases) is compared with another group known not to have the condition (controls) with respect to the presence of certain risk factors in the past.[3] They are efficient for the study of rare diseases. Cross-sectional studies, also known as prevalence studies are ideal for determining the status of a disease or condition in a single slice of time. A cohort study is an observational study in which subjects are sampled based on the presence or absence of a risk factor of interest and followed over a period of time for the development of the condition(s) of interest. Cohort studies can be either prospective or retrospective (historical cohort study). Case-series studies have traditionally been considered by biomedical scientists as the weakest study design and some would not consider them as studies at all. However, qualitative studies which include case studies, interviews and focus group research are now being accepted as central to research in family medicine which as a science lies somewhere between the biomedical and social sciences. General practice research designs range from the purely experimental and quantitative designs at one end to the

purely descriptive at the other. For example, a drug efficacy trial may employ an experimental design while a study of school phobia would better be studied using a qualitative design.

Selection of Subjects

Having noted the study design, the critical reader seeks information regarding the subjects (humans, animals, materials) involved in the study. The sampling method used is critical to the generalizability of the study. The author of the research report should provide information on how subjects were selected for the study and, if appropriate, how treatment assignments were made. The most generalizable sampling method is a random sampling where the determination of the treatment group assignment is based on probability and is not influenced by the patient's or researcher's preference. Ideally, the type of randomization process should be stated. If the selection was not randomised, are the subjects representative of the population? It is important to note whether the inclusion and exclusion criteria for selection of subjects are specified and how consent to participate in the study was obtained from them. If a control group was used, how was it chosen? Information on response rates, the number of subjects followed up and numbers lost to follow-up should be available though some authors prefer to state this in the 'Results' section.

Bias

The reader should ensure that systematic bias is minimised or avoided. Systematic bias is anything that erroneously influences the conclusions about groups and distorts comparisons between the groups.[4] In order to avoid systematic bias, groups of subjects being compared should be as similar as possible except for the particular difference being examined. Types of bias in clinical research include selection bias, information bias and confounding bias. Selection bias arises from the manner in which the subjects are selected for the study. In case-control studies, selection bias occurs when cases and/or controls are different from the populations they supposedly represent. Information bias occurs when inaccurate information is gathered about the study subjects. This can arise as a result of misclassification of cases or controls, inability to recall information or errors during clinical measurements. Confounding bias refers to a systematic error that occurs from the mixing of the effect of an exposure of interest with other associated correlates of the disease outcome. In a study relating industrial pollution to chronic obstructive airway disease, smoking becomes a confounder (confounding variable).

Measurement

The evaluation of measurements used in medical research is often a complex task. Information should be available to the reader on the measurements used to determine the outcome variable (e.g. disease) and the predictor variable(s) which include the risk factor(s) or exposure(s) under investigation. The reader should judge whether the correct instrument was used in the measurement and whether the researcher measured what should have been measured in relation to the research question. Is the measurement used valid and reliable? A valid instrument is one that measures what it sets out to measure. The reliability or repeatability of an instrument is the estent to which an instrument minimises error on repeated measurements i.e. repeated measurements in stable subjects gives similar results. Obviously, if the instrument used (e.g. questionnaire) is not valid or reliable, the results of the study would be open to question. Ideally, the researcher should provide a clear description of the instrument used, and discuss how they assess the validity and reliability of the measuring instrument. Information should be available on measurements performed in the control group.

Ethical Issues

Ideally, the author should state how ethical considerations and confidentiality of the participants were safeguarded. This should include information on how consent was obtained from subjects, the risks and benefits involved, their freedom to question any aspect of the study and the implications of refusal to participate and withdrawal from the study.

The Sample Size

While reading a clinical trial comparing a new therapy to an existing form of treatment, the reader has to take note of the sample size as to ensure that it is sufficiently large to detect clinically significant differences that have sufficient statistical power. The

power of a study refers to the probability of detecting a specified difference at a specified significance level, should a difference exist.' A power in the range of 80-90% is usually considered reasonable. There are important statistical and ethical implications in the choice of sample size for a study. Studies with a very small sample size may be unable to detect clinically important effects and also are an unnecessary utilisation of subjects and material resources. The approach to the calculation of sample size is complex and is beyond the scope of this article. However, there are a few specific problems that may arise with using an inadequate sample size.[5] First, there is a risk of falsely concluding that a significant difference exists in the outcome between two groups when in fact the finding has arisen by chance (alpha or Type 1 error). This is a false positive finding. A second possibility, is that the study may show no significant difference between the two groups when in fact a difference exists (beta or Type II error). In other words, the treatment may in fact be effective but the study shows that it is not (false negative finding). Other determinants or sample size include variability (standard deviation) which is the extent to which the subjects studied differ from each other in their characteristics and the magnitude of difference that the researcher aims to detect. Here the researcher needs to have a sample size that is large enough to detect clinically meaningful differences between the groups studied. Even so, the duration of the study must be of sufficient length for the intervention to have an influence on the outcome variable. However, in practice, the sample size is often influenced by other factors such as the availability of resources (manpower, funds, time) and prevalence of the disease. A clear and detailed explanation of the assumptions relating to the power of the study is a good indicator of a well-conducted study.

The Results

This section should present the data outcome of the study directed at questions stated in the introduction section. Charts, tables and graphs should be presented clearly and objectively and in sufficient detail to be reasonably intelligible without having to refer to the text (coherent on their own.) The reader should see that the numbers add properly and that the tables can be reconciled. The measurements made and the relevant outcomes should be clearly stated. The number of subjects lost to follow up or who withdrew from the study for various reasons and the statistical handling of data relating to this category of patients should be stated. Ideally, appropriate information should be available about the baseline measures of the group or groups studied e.g. age, sex, ethnicity, occupation, risk factors etc. If the groups were dissimilar on baseline measures, did the researcher perform appropriate analysis to account for their differences? This is important as 'significant differences' could result from the lack of comparability of the individual characteristics of the group.

Statistical Testing

The author should clearly state the reason for choosing a particular statistical test. If obscure statistical tests results are used, the author should justify their choice and provide a clear description of the test.[6] Statistical test results are usually given as either p-values or confidence limits. The actual results obtained should always be provided rather than just stating the values obtained by complex statistical calculations as undue emphasis on significance testing may result in the reader overlooking the actual magnitude of the difference between treatments or comparisons. This is essential as all that a significance test tells us is whether the results could have arisen by chance.[7] A p value of 0.05 only means that the result obtained would have arisen by chance on less than one in 20 occasions. The mention of 'confidence intervals' especially in clinical trial reports, provides the reader a range which at a certain confidence level (e.g. 95%), includes the real treatment differences. The 'confidence interval' provides a range of outcomes with which the results are compatible and is particularly important in situations where the result of a treatment comparison is stated as 'non-significant'. A useful index in interpreting the efficacy of a diagnostic or screening test is the 'likelihood ratio'. It expresses the odds that a given level of a diagnostic test would be expected in a person with the target disorder as opposed to one without the disorder.[8] In other words, the 'likelihood ratio' of a positive test is how much more likely is a positive test to be found in a person with a target disorder than in someone without it. If

statistically significant differences are mentioned, the reader has to consider their clinical and/or social significance as well.

Multiple Comparisons

Sometimes, a group of subjects are measured or compared repeatedly at different points in time resulting in the problem of 'multiple comparisons'. Ideally, there should be some discussion on the effect of multiple comparisons on the observed results. The more the significance tests that are performed in a clinical trial, the more likely are false positive results to appear by chance. [9] Therefore, multiple analyses increase the risk of false positive results. It is generally accepted that studies that report positive results are more likely to be published than those with negative results and thus the risk of false positive results in the literature may be high.[10]

The Discussion and Conclusions Sections

This is often an interesting and easier section to read compared to the result section. At this stage, the reader should see that the conclusions are consistent with the research questions stated in the 'Introduction' and the findings in the 'Results' sections. The researcher should identify similarities and differences between his findings with those of other researchers in the field. Discussion on drug trials should provide relevant information on the safety, tolerability, efficacy and price of new treatments comparing them with existing ones. Ideally, the results of the study should be within the context of existing knowledge. If marked conflicts exist with the findings of other researchers, the investigator should discuss the likely reasons for the differences. It is important to note whether the researcher has over-generalised his findings or tends to extrapolate beyond the data generated in the study.

Generalisability or external validity is the extent to which the results of a particular study can be used to make an inference about other populations. This is more often an issue for the critical reader to decide rather than for the researcher. Has the researcher addressed the limitations of his/her study and offered suggestions for further research on the subject? Finally, however convincing the study design, the results or the arguments put forth by the researcher, the critical reader should pose the final question - Am I going to believe and practice all that I have just read?

References

1. Dawson-Saunders B, Trapp RG. Basic and Clinical Biostatistics. Chapter 15, Reading the medical literature. Connecticut USA: Appleton and Lange 1990; 264-76.
2. Ibid.
3. Herman J, Slawson D. Case control studies. Fam. Med. 1990; 22(1):52-6.
4. Greenhalgh T. Assessing the methodical quality of published papers. In: How to read a paper: the basics of evidence based medicine. London: BMJ Publishing Group, 1997;305-8.
5. Gordon TM, Gordon DM. Clinical trials. Medicine International. Vol 5(18); 1992:4216.
6. Greenhalgh T. Statistics for the non-statistician. In:How to read a paper:the basics of evidence based medicine. London:BMJ Publishing Group, 1997; 315:364-7.
7. England JM. Medical Research, A statistical and epidemiological approach. Chapter 12, Hints on reading a publication. Churchill Livingstone 1975; 123-4.
8. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. The interpretation of diagnostic data. In: clinical epidemiology, A basic science for clinical medicine. Little, Brown and Co. Toronto, 1991:119.

9. Gordon TM, Gordon DM. Clinical trials. Medicine international. Vol 5 (18); 1992:4220.

10. Jewell D. Reading scientific articles or how to cope with the overload. The Practitioner. 1988;232:720-5.

M.C.Q.s on the Article on Critical Reading and Appraisal of Medical Literature

1. When evaluating the efficacy and effectiveness of a preventive intervention, the best study design to use is:

   a) Descriptive study design.
   b) Case-control study.
   c) Case report.
   d) Randomised controlled trial.
   e) Cohort study.

2. Which of the following statements regarding study designs is/are true?

   a) Cohort studies can be either prospective or retrospective in nature.
   b) The cohort approach is useful in calculating prevalence rates of diseases.
   c) The cohort study has the strongest observational design for studying a cause and effect relationship.
   d) In a cross-sectional study, the exposures to the conditions of interest are studied simultaneously.
   e) The case control design is efficient for the study of rare diseases.

3. In a clinical trial evaluating two antihypertensive medications, the aim of randomization is to:

   a) Select motivated patients for the trial.
   b) Obtain treatments groups of identical size.
   c) Enhance patient compliance with treatment.
   d) Eliminate the possibility of observer bias.
   e) Obtain study groups with comparable baseline characteristics.

4. In relation to sample size determination, which of the following statements are correct?

   a) Increasing the sample size reduces the chance of a Type I error.
   b) Increasing the sample size increases the chance of a Type II error.
   c) Increasing the sample size reduces the expected difference in outcome between the groups studied that is detectable.
   d) Sample size determination is influenced by the amount of variability in the subjects to be studied (standard deviation).
   e) Is determined based on the primary research end point.

5. The effectiveness of a new drug therapy for breast cancer is compared with what of a standard regimen in a limited clinical trial. No significant difference in 5 year survival rates were reported although the new drug was in fact superior. The failure to prove the greater effectiveness of the new drug may be due to:

   a) Measurement bias.
   b) Type II error.
   c) Type I error.
   d) Blinding.
   e) Placebo effect.