IJSPT

INVITED COMMENTARY

LEVELS OF EVIDENCE IN MEDICINE

Peter McNair, PhD, PT[1]
Gwyn Lewis, PhD[1]

ABSTRACT

Levels of evidence allow clinicians to appreciate the quality of a particular research paper quickly. The levels are generally set out in a hierarchical order, which is based largely upon the experimental design. While there are ideal designs for studies examining the effects of interventions, risk factors for a clinical condition or diagnostic testing, in most instances researchers have had to make compromises and these subsequently decrease the quality of their work. This paper provides information concerning how those compromises relate to subsequent levels that are given to a piece of research. It also provides an understanding of issues related to evaluating papers, and suggest ways in which the reader might discern how relevant a paper might be to one's clinical practice.

Key words: levels of evidence, research design, study quality

CORRESPONDING AUTHOR
Peter McNair
Health and Rehabilitation Research Institute
Faculty of Health and Environmental
Science
Auckland University of Technology
Private Bag 92006
Auckland
New Zealand
P: +64 9 9219999
F: +64 9 9219620
E: peter.mcnair@aut.ac.nz

[1] Health and Rehabilitation Research Institute, Auckland University of Technology, Auckland, New Zealand

INTRODUCTION

During the 1990s, the term evidence based medicine (EBM) became notably more apparent in research and clinical literature. As the name suggests, it referred to examining the research evidence for making clinical decisions, and as such it was more firmly grounded in the assessment of the science supporting clinical decision-making, rather than a reliance on the experiences and subjective perceptions of so called authorities or experts.[1] For EBM to have credibility, there needed to be a systematic manner in which clinical research was assessed, and this demanded the development of levels of evidence to ultimately appreciate and assess the quality of research available in answering a particular clinical question. Initially, efforts on assessment of quality were focused upon intervention studies, examining the degree of effectiveness of treatments for clinical disorders, however, in recent years such efforts have expanded to include other key clinical research areas such as diagnosis and risk factors. The purpose of this paper is to describe the key elements that determine the levels of evidence that subsequently allow the most appropriate or efficacious clinical decision to be made for the patient.

STUDY DESIGN HIERARCHIES PROVIDE AN INITIAL STARTING POINT

Physical Therapists are often interested in studies that involve treatment interventions, identifying risk factors for succumbing to an injury or disease, and diagnosis of clinical conditions. In each of these areas, there are a number of different study designs that can be implemented. These designs may dictate the potential importance of the studies findings in its field. The design that a researcher chooses should be that which most appropriately answers the question being posed.[2] However in many cases, it reflects the resources that researchers have at their disposal and the practicalities of undertaking the research. Resources required for studies may involve physical space and equipment, expertise in data collection, administrative processing of data, statisticians for analyzing data, and patient availability. In most cases, a researcher does not have the opportunity to cover all of these resources to the maximum level possible. Because of this, compromises are made and these often affect the choice of design to be utilised during the research process.

In studies concerning interventions, risk factors and diagnosis, the strength of an experimental paper's design is rated upon a scale that has 4-5 levels and may be regarded as a hierarchy with level 1 being the highest. In the current paper, the hierarchies presented are based on those recommended by the National Health and Medical Research Council of Australia.[3] However, there are others[4] and they generally follow the same pattern, being different only in the alphanumeric nomenclature given to the levels of the hierarchy (eg: 1a or IIa etc). While one design may be high in the hierarchy for a particular question to be answered, it may not fare so well for a different question. For instance, while a prospective cohort study may be very effective at identifying risk factors, such a design does not provide professionals with the best evidence of a treatment's effect on a particular clinical condition. For the latter, a randomised controlled trial (RCT) would be more appropriate. Thus, it is important to recognise that different study designs have particular features that may make them advantageous for answering a certain type of research question.

If possible, always look for systematic reviews when searching the literature. A Level 1 rating is reserved for a systematic review of the *experimental* papers. In such a paper, the quality of the designs and the findings of all the individual experimental papers are assessed in a systematic manner to provide an overall assessment or answer for a particular study question. However, it should be noted that not all systematic reviews automatically reach Level 1. If the papers that were reviewed were primarily of studies with poor designs, then the strength of evidence for the providing the answer to the question posed is lower, and the systematic review no matter how well it was conducted will not receive Level 1 status.[5] Thus, the experimental papers upon which the review is based should determine the validity and strength of the review's findings.

Even when a systematic review has utilised papers with the strongest possible designs, the professional needs to appreciate a number of other factors that will influence its importance. These include the number of papers that have been reported upon, and the consistency of the results across papers. One should also appreciate the degree to which the findings apply to the clinical population of interest and what the implications are in respect to applying them in clinical practice, that is, could they be reasonably implemented. On the above-mentioned scale, the highest quality *experimental* designs are rated with a Level 2 and lesser-rated designs receive Levels that decline to 4-5.

Interventions

For studies examining treatment interventions, randomised controlled trials (RCTs) provide Level II evidence, the strongest level of evidence below a systematic review. Not surprisingly, the two key criteria for these study designs are the incorporation of at least one control group and the randomisation of participants.[6] Without a control group, it is impossible to determine how participants would have changed over time without the experimental intervention. For instance, changes may have occurred due to disease progression or spontaneous recovery. The specific conclusions that can be drawn regarding the experimental intervention are critically dependent on what the control group receives. For example, researchers could compare the effects of icing on acute knee pain to a control group who received no specific intervention, or they could give the control group a bag of peas that are at room temperature to place over their knee for the same period of time. In the first example, the only conclusion that could be drawn is that icing is more effective at reducing pain than no treatment, whereas in the latter example, by controlling for effects associated with receiving a physical intervention to the knee and for the time of application, a researcher could therefore make more specific conclusions regarding the effects of ice itself. In terms of randomisation, the crucial criterion for a RCT is that neither the participant nor the experimenter should be able to predict which group the participant will be allocated to. Commonly accepted randomisation procedures include a coin toss, random number generator, drawing group allocation from an envelope. While researchers may design more complex procedures to ensure that group characteristics are matched on important factors and that participant numbers are balanced between groups, the final determination of group allocation for each participant should be due to chance alone.

One step down from an RCT is a pseudo-RCT, which provides Level III-1 evidence. In these study designs, there is still an appropriate control group but group allocation is not strictly randomised. Group allocation in pseudo-RCTs is dictated by a set rule such as date of birth or participant number. These are weaker randomisation procedures as the experimenter can have knowledge of the group to which a participant will be assigned. The ability to predict group allocation introduces bias into the study as this knowledge can affect the decision about whether to enter the participant into the trial, which may bias the results of the trial overall.

The next level of evidence, Level III-2, incorporates non-randomised controlled trials and two types of observational studies. Non-randomised controlled trials have marked group selection bias. For example, participants may allocate themselves into groups by choosing to receive a treatment, or participants presenting to a particular treatment provider might be always allocated to the experimental intervention and those that present to another treatment provider might receive a control intervention only. Observational designs include cohorts in which a group of people who are exposed to a particular intervention are followed over time and their health outcomes compared to a similar group of people who were not exposed to the intervention. Another example of an observational study is the case-control design, in which people with a selected condition are identified and their history of exposure to an intervention is compared to a similar group of people who do not have the condition. In all of these study designs, the researchers are not in control of group randomisation and thus the potential for selection bias is substantially higher than in RCTs. This selection bias means that there will be an inherent risk that confounding factors, or factors other than the intervention of interest, are influencing the results of the study. However, it is important to recognise that there are some research questions and interventions to which researchers cannot apply the principles of randomisation and have subjects assigned to different groups., e.g. abortion or obesity, or whether parachutes are an effective life saver. In such situations, the observational designs are the best or only alternative, and hence they can be extremely valuable.[7]

The final group of studies providing Level III evidence (Level III-3) are comparative studies with non-controlled designs. These are non-randomised studies where a group of people receiving the intervention of interest are compared with previous or historical information, or to another group receiving another intervention in another study. The key limitation of these studies is the lack of a concurrent control group, and thus it is not possible to determine the specific ***effects*** of the intervention in the population as there is not a suitable comparative group. The attempt to make up for the lack of a control group by comparing to historical data or other studies provides an improvement over non-comparative studies (see case series below), but is still limited. For example, comparison to historical data on disease progression may be confounded by changes in disease management, specific characteristics of the participants tested, or variations in the assessment of outcome measures.

The lowest level of evidence (Level IV) is provided by case series that have no comparison group. These are usually pre-test - post-test comparisons of outcomes following an intervention in a single group. Obviously, the lack of a control comparison severely limits the strength of the findings and the conclusions that could be drawn. These study designs will often incorporate the addition of a second pre-test measure following a baseline, control period. This control period and additional baseline measure marginally strengthen the design of the study by enabling participants to serve "as their own control". Case series study designs are commonly used for feasibility studies to demonstrate the potential efficacy, safety, or practicality of an intervention before implementation in a larger, more robust study.[8]

Risk factors

In the intervention section above, we described observational study designs such as the prospective cohort and the case control. While not the best choice of design for examining interventions where subjects can be randomised into groups, they can be very powerful in the study of risk factors associated with the development of clinical conditions.[9] In the aetiology hierarchy, the strongest of the observational studies is the prospective cohort receiving level II. As the name suggests, it follows a group of similar individuals (eg: forestry workers) over time to examine whether a particular factor (eg: vibration from chain saw use) influences the occurrence of an outcome (osteoarthritis in the hand). A key point is that the occurrence of the outcome has not occurred at the commencement of the study. Such a design allows a consistent measurement of exposure across all the study participants and consistent measurement of the criteria that determines the outcome (eg: the presence of osteoarthritis in the hand). Cohort designs can be prospective or retrospective with the latter being at a lower hierarchal level. The key difference is that the data related to the exposure and the outcome has already been collected in the retrospective design. In many instances, the risk factor and/or outcome of interest was not the reason for the original study.[10] For example, while a prospective study may have primarily been run to examine vibration levels as a risk factor for osteoarthritis of the hand in forestry workers, data might also have been collected on specific safety procedures and injuries that occurred in this cohort. Such data can be linked retrospectively and associations between variables can provide important findings. However, because the retrospective study was not the original intention, the same degree of standardisation of the data collection procedures and the precision in which they were collected is unlikely to have been undertaken and therefore the design is not as strong as a prospective study.

At the next level in the hierarchy of designs for examining risk factors is the case-control study. In this design two groups are identified, one that has a clinical condition of interest, and another that does not. For instance, a group of forestry workers with osteoarthritis of the hand would be the case group and they would be compared to a group of forestry workers without osteoarthritis of the hand. That comparison might involve examining potential physical risk factors, (e.g. tools used, tasks performed, times and volume of work) that were undertaken by both groups over a specified time to highlight a risk factor or set of factors that are different across the groups. This design is weaker than the cohort design as only the outcome (osteoarthritis of the hand) has the potential to have been measured in a standardised and precise manner.[10] Even then, one of the most notable criticisms of this design is that the criteria for being included in either the control or case groups may be insufficient to accurately represent those of the wider population with and without the condition of interest.[9] This is particularly so, when the case-control design is targeting risk factors for a rare condition. Characterising risk factors associated with rare conditions is a key strength of the case control. The alternative, if one were to use a prospective cohort, means waiting for sufficient cases to contract a disease so that its risk factors might be characterised well, and that may never eventuate.

Cross sectional study designs and case series form the lowest level of the aetiology hierarchy. In the cross sectional design, data concerning each subject is often recorded at one point in time. For instance, a questionnaire might be sent to a district where forestry is a predominant industry. It might ask about the presence of osteoarthritis in the hand. In doing so, the prevalence of the disorder can be established. Some information related to exposure might also be collected and associations might be observed, but it is difficult to be confident in the validity of these associations. Thus, information gained from the cross-sectional study is often a starting point that provides the impetus to use a more powerful design to substantiate the initial findings.

Diagnosis

For diagnostic studies, the basic design utilized is very similar across most studies, and the higher levels of the hierarchy are based on meeting specific methodological criteria within that design. To receive Level II strength, the design is usually a prospective cohort, and the comparison it makes between a diagnostic test and a reference standard requires the following criteria:[11] All subjects should receive the reference standard, and that standard should be the best evidence available for determining whether the condition of interest is present. For studies, involving primary care, this will often be a scanning or electrophysiological procedure and might also include an anaesthetic block, while in studies involving tertiary care patients, the reference standard is often what is observed at surgery. The diagnostic test and the reference standard should also be completely independent of one another. It is crucial that the reference standard and the diagnostic tests are clearly described so that others can replicate them. The persons performing the diagnostic tests on the patients should not have knowledge of the results of the reference standard and similarly those performing the reference standard should have no knowledge of the results of the diagnostic test. The patients participating in the study must be well described, and represent those with mild as well as severe levels of the condition of interest who are recruited in a consecutive manner, and at the end of the study they are all accounted for.

Studies where the subjects are not consecutively recruited are assigned level III-1 strength. When the criteria relating to reference standards are partially compromised, a study is regarded as level III-2. When a study uses a group of subjects that don't include a wide spectrum of those likely to have the condition, or don't identify specific potential subgroupings that might affect the results, it is assigned level III-3. Such studies are often case-control designs where there are narrow criteria for inclusion in either the case or control groups, which can ultimately affect the generalizability of the results.[12] The lowest level (IV) is reserved for those studies that lack a reference standard.

IRRESPECTIVE OF DESIGN, THE QUALITY OF STUDIES IS IMPORTANT

While hierarchies provide the professional with a guide to how well a study design might answer a question, one must also consider how well that design has been implemented.[5] Within each design, there is a set of criteria that should be subscribed to, to make the design as robust as possible. The RCT may be at level II on the design hierarchy, and hence a good choice of design for studies examining the effects of an intervention. However, if that RCT has insufficient subject numbers to detect a reasonable difference across groups or blinding of subjects was not undertaken, or there were notable dropouts, then one should question the value of the results from that study, despite the design being the most appropriate. A study with a design lower on the hierarchy that has been undertaken well may provide more valid information.

There are numerous scales or checklists to choose from within the literature to assess the quality of individual research studies across the domains of interventions, aetiology, and diagnosis. The key sources of bias that might threaten the validity of the results of studies generally relates to the selection of patients, randomization, therapeutic regime, withdrawals, blinding, and statistical analyses.[6] Be aware that some checklists are extremely extensive[13] and include questions on issues that may not actually have the potential to bias the results, which is the primary reason for your assessment of the methodological quality.

The answers to checklist questions concerning methodological issues may be categorical (eg: bias present or not) or may be graded (e.g. 1 to 4). In some instances, the answers are weighted according to how important the checklist developer thought the bias might affect the results. Generally, the weightings of checklist questions have been subjectively applied with little if any empirical support, and subsequently total scores across checklists can be quite different.[6] Where weighting has not been applied across questions, the assumption is that all issues are of the same value and that is arguably not so. In light of these potential issues, at the Cochrane Collaboration Higgins et al[14] have indicated that readers refrain from giving an overall score to a paper on its methodological quality, but rather to identify whether methodological quality criteria have been met or not met, and in the latter case, how relevant the issue might be to the size of the effects observed in the study. This strategy makes it much harder for an individual to discern whether a particular paper is one that should be given more or less consideration, in respect to clinical decisions to be made. If clinicians are expected to assess the merits of individual experimental papers, this is an area that must be addressed further for more types of studies. Key sources of questionnaires for assessing the quality of intervention, risk factor and diagnostic studies are provided by Higgins et al,[14] Hayden et al,[15] Bossuyt et al,[16] and Whiting et al,[17] respectively.

APPLYING WHAT IS FOUND IN THE LITERATURE TO CLINICAL SCENARIOS

Assuming that papers have been identified that perform well from a methodological perspective, and their designs are well placed on the hierarchy for answering a particular question, finding papers that include participants who are similar to the patient(s) of interest to the professional is important. Such consideration should include an assessment of the level of severity of the groups under study (eg: mildly, moderately or severely affected), together with the amount of treatment they were being given, and the timing of that treatment within their disease/injury healing process. Furthermore, check when the researchers made their assessments to determine change in the participant's status. Ask whether these are realistic time points to do an assessment, and if the follow up was appropriate to determine the longer-term effects.

It is also important that clinicians look beyond the treatment effect of an intervention to get a balanced view of its merits. Consideration should be made not only of the benefits but also the potential harm associated with a particular treatment. For instance, a new regime for treating acute muscle tears might be developed and shown in a well-conducted RCT to allow players to return to sports much earlier than anything currently available. However, that same regime may induce side effects, perhaps a greater likelihood of the injury recurring 6-12 months later due to the laying down of excessive scar tissue in the early stages of the rehabilitation regime. Examination of such points will allow the professional to make a better judgement concerning the relevance of the papers to the clinical decision at hand.

GUIDELINES PROVIDE A SYSTEMATIC REVIEW AND A SET OF RECOMMENDATIONS

Based on the information presented above, it would seem a monumental task for therapists to assess a series of individual papers and thereafter make an informed decision concerning every clinical problem that they face, particularly those where the patient is atypical, and does not resemble the subjects presented in studies. To make the task easier, guidelines have been developed to answer specific clinical problems/questions and provide recommendations. Because of the resources required, guidelines are usually initiated by organisations such as specialist groups in a field of medicine/allied health or a national health agency. These organisations convene a guidelines panel that is usually composed of scientists, clinical specialists, statisticians, patients and lay people, and they are supported by data analysts and administrators. Their first step is to identify the question of interest and the key outcomes associated with that question. They then assess systematic reviews (Level 1 evidence in the hierarchy) that have been previously published or specifically undertake their own systematic review. In doing so, they provide a summary of the quality of the research undertaken, the consistency of the results across studies, the magnitudes of the intervention's effects observed in patient subgroups, the benefits versus the potential harm associated with a treatment, and whether the health benefits of a treatment are worth the costs of providing them.[18] Most importantly though, guidelines include recommendations and these are often quite definitive, being categorised as 'strong' or 'weak'. Guyatt et al[19] describe these as reflecting a trade off between the benefits of treatment against the burdens of receiving it together with its risks; while taking into account the accuracy and strength of the data supporting the intervention. If the data analysed from experimental papers indicates that an intervention has a large effect and the risks and burdens associated with the treatment are low, then a strong recommendation can be made to implement it. Where there are inconsistencies in findings or small treatment effects or notable risks, the recommendation for the treatment/intervention might be regarded as 'weak', and the patient's particular circumstances may then play a greater role in whether a particular treatment is implemented.

Given the extent and thoroughness behind the construction of guidelines and the inclusion of recommendations, they are an important source for guiding clinical decision making and should be searched for early in your examination of the literature.

THINK BEYOND THE SCIENCE

While the current paper has focused upon the quantitative assessment of evidence, it cannot be regarded as the sole means by which professionals make clinical decisions. It is important that therapists continue to appreciate the individuality of each patient and the personal circumstances that they bring with their pathophysiological issues. While at present, qualitative research does not have a formal place in levels of evidence, there is without doubt evidence for its importance in providing insights into patients' viewpoints on how the clinical condition and its treatment has influenced the lives that they lead.

Therefore, professionals must continue to value highly how we interact and react to each patient's situation, continually striving to be effective listeners and communicators, as well as being advocates of the best research evidence to help all patients improve the quality of their lives.

Table 1. *Key Points summary.*

Key Points Summary:

- ❖ Different types of questions require different research designs.
- ❖ Look for clinical guidelines and systematic reviews first.
- ❖ Consider the hierarchy of designs within diagnostic, intervention, and risk factor studies.
- ❖ Consider the *quality* of papers irrespective of the design.
- ❖ Assess how well the subjects recruited in studies match your patient's characteristics.
- ❖ Be aware of not only the benefits of treatments but also their potential risks.
- ❖ Don't forget to assess the patient's circumstances, and how it might influence your evidence-based clinical recommendations to them.

REFERENCES

1. Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, Wilson MC, Richardson WS. 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;284:1290-6.
2. Sackett DL, Wennberg JE. Choosing the best research design for each question. *BMJ.* 1997;315:1636.
3. NHMRC. *NHMRC additional levels of evidence and grades for recommendations for developers of guidelines.* Available at www.nhmrc.gov.au/guidelines/publications
4. Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, Dawes M. *Evidence-based medicine levels of evidence.* Oxford Centre for Evidence-Based Medicine; 2001.
5. Glasziou P, Vandenbroucke JP, Chalmers I. Assessing the quality of research. *BMJ.* 2004;328:39-41.
6. Hadorn DC, Baker D, Hodges JS, Hicks N. Rating the quality of evidence for clinical practice guidelines. *J Clin Epidemiol.* 1996;49:749-54.
7. Vandenbroucke JP. Observational research and evidence-based medicine: What should we teach young physicians? *J Clin Epidemiol.* 1998;51:467-72.
8. Vandenbroucke JP. In defense of case reports and case series. *Ann Intern Med.* 2001;134:330-4.
9. Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet.* 2006;7:812-20.
10. Mann CJ. Observational research methods. Research design ii: Cohort, cross sectional, and case-control studies. *Emerg Med J.* 2003;20:54-60.

11. Fritz JM, Wainner RS. Examining diagnostic tests: An evidence-based perspective. *Phys Ther.* 2001;81: 1546-64.

12. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299:926-30.

13. Chalmers TC, Smith H, Jr., Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. *Control Clin Trials*. 1981;2:31-49.

14. Higgins J, Altman D, Sterne J. Chapter 8: *Assessing risk of bias in included studies.* In: Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions: The Cochrane Collaboration; 2009.

15. Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. *Ann Intern Med*. 2006;144:427-37.

16. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: The stard initiative. *Ann Intern Med*. 2003;138:40-4.

17. Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J. Evaluation of quadas, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol*. 2006;6:9.

18. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Varonen H, Vist GE, Williams JW, Jr., Zaza S. Grading quality of evidence and strength of recommendations. *BMJ.* 2004;328:1490.

19. Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B, Raskob G, Lewis SZ, Schunemann H. Grading strength of recommendations and quality of evidence in clinical guidelines: Report from an american college of chest physicians task force. *Chest.* 2006;129:174-81.