Journal of Clinical Nursing 2003; **12**: 77–84

Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions

DAVID EVANS BN, MNS, PhD, RN, DipN
Lecturer, Department of Clinical Nursing, University of Adelaide, South Australia 5005

Accepted for publication 7 March 2002

Summary

• A number of hierarchies of evidence have been developed to enable different research methods to be ranked according to the validity of their findings. However, most have focused on evaluation of the effectiveness of interventions. When the evaluation of healthcare addresses its appropriateness or feasibility, then existing hierarchies are inadequate.

• This paper reports the development of a hierarchy for ranking of evidence evaluating healthcare interventions. The aims of this hierarchy are twofold. Firstly, it is to provide a means by which the evidence from a range of methodologically different types of research can be graded. Secondly, it is to provide a logical framework that can be used during the development of systematic review protocols to help determine the study designs which can contribute valid evidence when the evaluation extends beyond effectiveness.

• The proposed hierarchy was developed based on a review of literature, investigation of existing hierarchies and examination of the strengths and limitations of different research methods.

• The proposed hierarchy of evidence focuses on three dimensions of the evaluation: effectiveness, appropriateness and feasibility. Research that can contribute valid evidence to each is suggested. To address the varying strengths of different research designs, four levels of evidence are proposed: excellent, good, fair and poor.

• The strength of the proposed hierarchy is that it acknowledges the valid contribution of evidence generated by a range of different types of research. However, hierarchies only provide a guide to the strength of the available evidence and other issues such as the quality of research also have an important influence.

Keywords: evidence, hierarchy, levels of evidence, research.

Introduction

The past two decades have seen a growing emphasis on basing healthcare decisions on the best available evidence. This evidence encompasses all facets of healthcare, and

Correspondence to: David Evans, Lecturer, Department of Clinical Nursing, University of Adelaide, South Australia 5005 (tel.: +61 8 8303 6288)

includes decisions related to the care of an individual, an organization or at the policy level. Attention has also focused on the quality of the scientific basis of healthcare and, with this, recognition that not all evidence is equal in terms of its validity.

To aid the interpretation and evaluation of research findings, hierarchies of evidence have been developed which rank research according to its validity. The major focus of these hierarchies has been effectiveness and, as a result, the randomized controlled trial (RCT) has been commonly viewed as providing the highest level of evidence. While many valid approaches to research exist, they are often ranked at a level lower than the RCT although each approach provides its own unique perspective.

To address this, a hierarchy for ranking research evaluating healthcare interventions was developed. This hierarchy differs from existing models in that it recognizes the contribution of evidence generated by a range of research methodologies.

Grading the evidence

It has long been recognized that not all research designs are equal in terms of the risk of error and bias in their results. When seeking answers to specific questions, some research methods provide better evidence than that provided by other methods. That is, the validity of the results of research varies as a consequence of the different methods used. For example when evaluating the effectiveness of an intervention, the RCT is considered to provide the most reliable evidence (Muir Gray, 1997; Mulrow & Oxman, 1997; Sackett *et al*., 1997). It is considered the most reliable evidence because the processes used during the conduct of an RCT minimize the risk of confounding factors influencing the results. As a result of this, the findings generated by RCTs are likely to be closer to the true effect than the findings generated by other research methods.

This confidence in the findings of research has important implications for those developing practice guidelines and clinical recommendations, or implementing the results of research in their area of practice. The aim during this development and implementation is to use the best available evidence. The problem that arises from this situation is how to determine the best evidence. To address this, hierarchies of evidence have been developed to allow research-based recommendations to be graded. These hierarchies or levels are used to grade primary studies according to their design, and so reflect the degree to which different study designs are susceptible to bias [National Health Service (NHS) Centre for Reviews and Dissemination, 1996]. Ranking research designs according to their internal validity not only grades the strength of the evidence, but also indicates the confidence the end-user can have in the findings.

Hierarchies of evidence

Hierarchies of evidence were first popularized by the Canadian Task Force on the Periodic Health Examination in the late 1979, and since that time many different hierarchies have been developed and used (Canadian Task Force on the Periodic Health Examination, 1979; Sackett, 1986; Woolf *et al*., 1990; Cook *et al*., 1992, 1995; Guyatt *et al*., 1995; Wilson *et al*., 1995). Until recently, these focused on effectiveness, and for this reason the RCT was most commonly listed as providing the highest level of evidence.

These hierarchies have used a range of different approaches to grading research. For example, one hierarchy for clinical recommendations used levels A1 through to C2 (Guyatt *et al*., 1995). Level A1 represented RCTs with no heterogeneity and a confidence interval (CI) all on one side of the threshold number needed to treat (NNT). Level C2 at the other end of this scale was assigned to observational studies with a CI overlapping the threshold NNT. NNT is the number of patients who have to be treated to prevent one event occurring (see Information Point in Vol. 10, no. 6, p. 783). Another hierarchy used a scale of level I through to level IV [National Health and Medical Research Council (NHMRC), 1995]. Level I was assigned to evidence obtained from a systematic review of all relevant randomized controlled trials, while level IV comprised opinions of respected authorities, descriptive studies, or reports from expert committees. The Cochrane Collaboration ranks the validity of studies on a scale of A to C, with A indicating that the study met all quality criteria (Mulrow & Oxman, 1997). One hierarchy that was used during the development of clinical guidelines used an alpha-numerical approach to rank both evidence and recommendations (Meltzer *et al*., 1998; Sackett, 1986). The highest ranking in this hierarchy was 'Grade A Recommendations supported by Level I evidence' (Cook *et al*., 1992).

With the increasing popularity of systematic reviews, these are starting to replace the RCT as the best source of evidence (NHMRC, 1995). More recently, one hierarchy listed N of 1 randomized trials as the highest level of evidence (Guyatt *et al*., 2000). N of 1 randomized trials use a single patient who is randomly allocated to the treatment and comparison interventions. Hierarchies have

now been developed to address a range of other areas, including prevention, diagnosis, prognosis, harm and economic analysis (Carruthers *et al*., 1993; Ball *et al*., 1998; Meltzer *et al*., 1998).

Ultimately, these hierarchies aim to provide a simple way to communicate a complex array of evidence generated by a variety of research methods. From the perspective of healthcare decision-makers, they provide a measure of the trust that can be placed in the recommendations, or alert the user when caution is required. However, the exact format and order of rank for research designs within these hierarchies have not been determined and existing systems have used a range of different approaches.

Determining best evidence

A limitation of current hierarchies is that most focus solely on effectiveness. Effectiveness is concerned with whether an intervention works as intended. While this is obviously vital, the scope of any evaluation should be broader. For example, it is also important to know whether the intervention is appropriate for its recipient. From this perspective, the evidence on appropriateness concerns the psychosocial aspects of the intervention and so would address questions related to its impact on a person, its acceptability, and whether it would be used by the consumer. A third dimension of evidence relates to its feasibility, and so involves issues concerning the impact it would have on an organization or provider, and the resources required to ensure its successful implementation. Feasibility encompasses the broader environmental issues related to implementation, cost and practice change.

Evidence on effectiveness, appropriateness and feasibility provides a sounder base for evaluating healthcare interventions, in that it acknowledges the many factors that can have an impact on success. This highlights the range of dimensions that evidence should address before healthcare interventions can be adequately appraised. It also means that, no matter how effective an intervention is, if it cannot be adequately implemented, or is unacceptable to the consumer, its value is questionable. The risk with available hierarchies is that, because of their single focus on effectiveness, research methods that generate valid information on the appropriateness or feasibility of an intervention may be seen to produce lower level evidence.

In response to these limitations of existing frameworks, a new hierarchy of evidence was developed that acknowledges the legitimate contribution of a range of research methodologies for evaluating healthcare interventions (see Fig. 1). This approach addresses the multidimensional nature of evidence and accepts that valid evidence can be generated by a range of different types of research. It does not attempt to diminish the value of RCTs, or the importance of determining effectiveness; rather, it accepts that RCTs answer only some of the questions. Importantly, this framework acknowledges the contribution of interpretive and observational research.

From a slightly different perspective, the hierarchy was also developed to serve as a framework during the production of systematic review protocols. In this context, the aim of the hierarchy was to help formulate review questions and to assist in determining what research could provide valid evidence when questions extended beyond the effectiveness of an intervention.

	Effectiveness	Appropriateness	Feasibility
Excellent	• Systematic review • Multi-centre studies	• Systematic review • Multi-centre studies	• Systematic review • Multi-centre studies
Good	• RCT • Observational studies	• RCT • Observational studies • Interpretive studies	• RCT • Observational studies • Interpretive studies
Fair	• Uncontrolled trials with dramatic results • Before and after studies • Non-randomized controlled trials	• Descriptive studies • Focus groups	• Descriptive studies • Action research • Before and after studies • Focus groups
Poor	• Descriptive studies • Case studies • Expert opinion • Studies of poor methodological quality	• Expert opinion • Case studies • Studies of poor methodological quality	• Expert opinion • Case studies • Studies of poor methodological quality

Figure 1 Hierarchy of evidence: ranking of research evidence evaluating health care interventions.

EFFECTIVENESS

Effectiveness has been the most common concern of systematic reviews and clinical guidelines. Effectiveness relates to whether the intervention achieves the intended outcomes and so is concerned with issues such as:

- Does the intervention work?
- What are the benefits and harm?
- Who will benefit from its use?

It can be argued that multicentre RCTs provide the best evidence for the effectiveness of an intervention because the results have been generated from a range of different populations, settings and circumstances (see Fig. 1). The findings from systematic reviews are generated in a similar manner, and so also provide rigorous evidence (Mulrow, 1987; Cook *et al*., 1998). As a result, the robustness and generalizability of evidence from both these approaches are better than what is generated by other research designs. This means that for the evaluation of effectiveness, the best evidence would be that produced by either of these approaches.

However, this is not the only source of good-quality evidence. A well-conducted single-centre RCT also produces results that are at low risk of error or bias, and so provides valid evidence on the effectiveness of an intervention. However, this evidence is ranked at a lower level because the findings are based on a single population. This means that factors unique to the study site, such as skill mix, available resources, staffing levels or expertise, may have an impact on the findings of the RCT.

For observational studies, such as case control or cohort studies, their place within the hierarchy of research designs is less clear and they have often been viewed as being at greater risk of systematic error than RCTs (Chalmers *et al*., 1983; Colditz *et al*., 1989; Miller *et al*., 1989). The concern with these studies is that they can distort the treatment effects, making them appear smaller or larger than they really are (Mulrow & Oxman, 1997). Recently, however, comparisons of the results of observational studies and RCTs evaluating the same intervention have questioned this claim (Benson & Hartz, 2000; Concato *et al*., 2000), and suggest that the findings of observational studies are similar to those produced by RCTs.

There are important differences between the RCT and observational study relating to their internal and external validity. Internal validity in this context is a measure of how easily differences in outcomes between comparison groups can be attributed to the intervention (Elwood, 1998). External validity refers to the way in which the results of a study can be generalized to the wider population (Elwood, 1998). The RCT minimizes the risks posed by confounding variables through processes such as randomization and strict inclusion criteria and, as a result, the RCT has high internal validity. However, because of these very processes, only a narrow spectrum of patients may qualify for inclusion in the study. This means that the external validity is low and so the generalizability of the findings of the RCT may be limited. Conversely, observational studies observe what is happening in practice and thus have a lower internal validity as a result of potential differences between comparison groups. As a result it is harder to attribute the differences in the outcome to the intervention. However, this lack of control means that observational studies are more firmly based in the real world, in that the comparison groups more closely reflect clinical practice. Therefore, it can be argued that observational studies have a higher external validity than RCTs. To put it more simply, gains in the internal validity of the RCT are achieved at the expense of external validity, while the high external validity of the observational study is achieved at the expense of internal validity.

In some situations, observational studies may be more suitable than the RCT, such as when measuring infrequent adverse outcomes, evaluating interventions designed to prevent rare events or those evaluating long-term outcomes (Black, 1996). Legal or ethical issues may also prevent the conduct of RCTs. Observational studies may also be the only option where clinicians or patients are unwilling to accept randomization as the mechanism for assignment of treatment (Horwitz *et al*., 1990). For some treatments, a sustained effort is required from the recipient and so their evaluation may require a different approach from the RCT (Brewin & Bradley, 1989). Additionally, these studies cost less than RCTs and allow evaluation of a broader range of participants (Feinstein, 1989). Finally, situations in which the results of RCTs contradict consistent findings from observational studies serve to highlight the need for caution (Guyatt *et al*., 2000).

From this perspective, it can be argued that both the RCT and observational study can contribute valid evidence related to the effectiveness of an intervention and therefore should have a role in any evaluation. The important difference between methods is that the RCT solely evaluates the intervention, while the observational study measures the intervention in clinical practice. When differences in results exist, they cannot be assumed to be solely due to the presence or lack of randomization (McKee *et al*., 1999). Factors such as differences in study populations, characteristics of the intervention or patient preferences may be responsible for the difference in

findings (McKee *et al*., 1999). However, these approaches can provide complementary evidence, and end-users must be aware that both methods have their strengths and weaknesses (McKee *et al*., 1999).

In addition to the studies already discussed, evidence is also produced by other methods such as non-randomized controlled trials, un-controlled trials, and studies with historical controls; however, their results are at greater risk of error (Dawson-Saunders & Trapp, 1994). With quasi-experimental designs, such as the non-randomized controlled trial, it is more difficult to show that any difference in outcome is the result of the intervention rather than differences between groups (Elwood, 1998). These non-randomized studies differ from observational studies because the allocation to comparison groups is made by the researcher rather than healthcare workers who are independent of the study. As a result of these factors, the risk of error or bias is high.

Uncontrolled trials may also be used to evaluate an intervention, but the lack of any comparison group makes interpretation of findings difficult. The exception to this is studies with dramatic results, for example, the administration of oxygen to a hypoxic person or adrenaline to a person in shock. However, for most situations, the evidence generated by uncontrolled trials should be regarded with suspicion, and must also be ranked at a lower level than the findings of RCTs or observational studies.

Finally, evidence about the effectiveness of an intervention may be generated through descriptive studies, expert opinion, case studies or poorly conducted studies. This evidence is at the greatest risk of error and is inadequate for evaluating the effectiveness of an intervention. As a result these methods are ranked as the lowest level of evidence.

APPROPRIATENESS

Appropriateness, in this context, addresses the impact of the intervention from the perspective of its recipient. It also relates to the impact of illness to enable this information to be integrated into healthcare management and to assist in the prioritization of care. Appropriateness is concerned more with the psychosocial aspects of care than with the physiological and, with regard to the intervention, is reflected in questions such as:

- What is the experience of the consumer?
- What health issues are important to the consumer?
- Does the consumer view the outcomes as beneficial?

The range of research methods that can contribute valid evidence on the appropriateness of an intervention is broader than that addressing effectiveness (see Fig. 1). Firstly, as with effectiveness, results generated by multicentre studies and systematic reviews represent the best evidence on the appropriateness of an intervention. However, the systematic review and multicentre study need not be limited to RCTs, but would focus on all methods that can reasonably be used to evaluate the intervention from the perspective of appropriateness. Recommendations based on these sources of evidence would be at least risk of error.

Good evidence can also be generated by a range of other research methods. As with effectiveness, a well-conducted single-centre RCT or observational study can provide valid evidence about the appropriateness of an intervention through a focus on psychosocial outcome measures. As previously stated, while experimental and observational studies evaluate the intervention from different perspectives, the evidence is complementary.

Interpretive studies can also contribute valid evidence, in that they represent the consumer's perspective on the treatment, illness or other such phenomenon, and thus help capture the subjective human experience that is often excluded from experimental research. This interpretive inquiry helps healthcare workers gain an understanding of everyday situations and experiences (Van Manen, 1990; Van der Zalm, 2000). While this information differs considerably from that generated by experimental or observational research, it contributes to our understanding of the impact of healthcare and is no less valid than that produced by other methods.

Evidence on appropriateness can also be generated by descriptive studies such as surveys, questionnaires and case studies. These contribute descriptive data related to interventions, their use and consumer responses. In addition to this, focus groups have emerged as a method for gathering information on the feelings and opinions of small groups of people, and so can aid in the evaluation of healthcare programmes (Beaudin & Pelletier, 1996; Robinson, 1999). This information offers another perspective on appropriateness and is valid evidence. However, its strength is less than that of the evidence produced by experimental, observational or interpretive research. Finally, evidence can also be generated by expert opinion or poor quality studies; however, this is at the greatest risk of error and as a result is ranked as the lowest level of evidence.

FEASIBILITY

Feasibility addresses the broader environment in which the intervention is situated and involves determining whether the intervention can and should be implemented. This focus acknowledges that the process of intentional

change in large organizations is very complex. In this context, feasibility is reflected in questions such as:

- What resources are required for the intervention to be successfully implemented?
- Will it be accepted and used by healthcare workers?
- How should it be implemented?
- What are the economic implications of using the intervention?

A broad range of research methods can reasonably be used to evaluate feasibility, and while each has a different focus, all offer important evidence (see Fig. 1). Once again, results generated by multicentre studies and systematic reviews can be considered the best evidence for evaluating the feasibility of an intervention. These reviews and studies need not be limited to synthesizing the findings of RCTs, but may focus on all methods that can reasonably be used to evaluate the intervention from the perspective of feasibility.

A well-conducted single-centre RCT can provide good evidence on the feasibility of an intervention. From this perspective, the RCT would be likely to focus on organization, utilization or implementation outcome measures or on activities that support the intervention, such as education programmes. Both observational and interpretive studies can generate valid evidence and would focus on issues related to implementation, acceptance, long-term benefits, or the impact of the organizational culture on implementation.

Other methods can also provide useful evidence on feasibility. For example, action research is able to explore the relationships between attitudes and specific aspects of care, to identify barriers to practice change, and to systematically develop knowledge related to practice (Meyer, 2000). As a result of this, action research can contribute legitimate evidence on which to influence and shape clinical practice. As with appropriateness, focus groups can also gather valid information from small groups of people (Basche, 1987; Beaudin & Pelletier, 1996), and so assist in evaluating healthcare programmes (Robinson, 1999). From the perspective of feasibility, this information would relate to such things as implementation, identifying barriers or determining what support is required. Descriptive studies can also provide information related to the feasibility of an intervention. However, the evidence generated by these methods would be ranked at a lower level than that produced from experimental, observational and interpretive research.

Finally, as with both effectiveness and appropriateness, evidence can be based on expert opinion, case studies or poor-quality research. However, this evidence is at the greatest risk of error and so is ranked at the lowest level of hierarchy.

Levels of evidence

The primary purpose of developing this hierarchy was to provide an indication of the validity and trustworthiness of different types of research. This process assists in the selection of the best evidence to guide clinical practice. However, each level proposed in this hierarchy differs from others, as described below.

- *Excellent:* This level of evidence provides the strongest scientific base for clinical practice. As this evidence is at the least risk of error, it is optimal for the development of practice guidelines and clinical recommendations.
- *Good:* This level of evidence also provides a sound basis for clinical practice and is at low risk of error. However, as it may have been generated by single studies, it also highlights areas where replication of research is needed.
- *Fair:* As this level of evidence will be at varying degrees of risk of error, it does not provide a strong evidence-base for clinical practice. However, these studies represent initial exploration of interventions and so assist in prioritizing the research agenda. The rationale for this is that while the evidence is at greater risk of error than the previous levels, it allows identification of potentially beneficial interventions that require additional investigation and evaluation.
- *Poor:* This level of evidence provides a poor basis for clinical practice and is at serious risk of error or bias. Additionally, while this evidence can help in determining research priorities, because there is a greater risk that it may be wrong, and therefore misleading, it is ranked below other forms of evidence.

Value of this approach

The benefit of this approach for grading evidence evaluating interventions is that it moves beyond having a single focus on RCTs. This broader focus is important because an RCT is unlikely to be able to answer all the questions needed for a complete evaluation. From this perspective, it acknowledges that, when evaluating an intervention, a variety of research methods can contribute valid evidence. This hierarchy also recognizes the greater strength of evidence when it has been generated from multiple populations, settings and circumstances. For this reason, evidence generated by properly conducted systematic reviews or multicentre studies should be considered the strongest evidence.

This approach to ranking evidence also legitimizes the perspective of the consumer of the intervention and so recognizes the pivotal role this should have in healthcare decisions. It also acknowledges the importance of the

psychosocial impact of interventions and that consumers' priorities on important health needs may differ from those of the providers of care. While the views of the consumer have long been part of the rhetoric, to date they have fitted poorly within the evidence-based framework. Through the use of this hierarchy, evidence addressing this aspect of the evaluation of an intervention can be ranked at a more appropriate level.

While an intervention may be effective, it must also be feasible to implement. This relates to such things as cost, healthcare workers' acceptance and the resources which will be required to support the intervention. This hierarchy recognizes that evidence addressing the feasibility of an intervention is as important as that addressing effectiveness. A broader approach to the ranking of evidence will provide a more robust scientific base for healthcare, in that it moves beyond the single focus of effectiveness that has dominated the evidence-based healthcare movement since its inception.

The gold standard

In the context of this hierarchy it can be argued that there are two interpretations of the label 'gold standard'. The common use of this term refers to the optimal research design to answer a question. Over the past decade, this label has most commonly been applied to RCTs evaluating the effectiveness of interventions. However, for research questions addressing issues other than effectiveness, different methods will be needed. The optimal research method will be determined by the type of question, and it is the method that produces the most valid evidence that should become the standard to which others are compared.

Secondly, the use of this hierarchical structure for grading evidence provides another interpretation of what is meant by the gold standard. The concept of gold standard could move beyond research design and refer to evidence addressing all three dimensions of the evaluation of an intervention. That is, evidence demonstrates that the intervention works, can be implemented and fulfils the needs of its consumers. Only when all these dimensions have been subjected to investigation can an intervention be fully appraised and the evidence considered to be of a gold standard.

Cautionary note

It must be acknowledged that the use of any hierarchy is, at best, a guide rather than a set of inflexible rules. A hierarchy provides the end-user of research with a framework to judge the strength of available evidence. Other issues, such as what outcome measures were used and the populations studied, also exert a major influence on the usability of the evidence. While I have used this hierarchy to provide a logical framework for a review, it has not been subject to any formal evaluation and so caution is needed. Finally, and most importantly, hierarchies cannot be used to rank evidence without some consideration of the quality of research. Regardless of the research method, if the processes used during the study were poor, then the findings must be regarded with suspicion.

Conclusion

The proposed hierarchy of evidence provides a tool by which research addressing the many dimensions of an intervention can be ranked at an appropriate level. This approach takes the emphasis away from the RCT, to one that accepts that different research designs may be required for different clinical questions. The focus on effectiveness, appropriateness and feasibility provides a broader base for evaluating healthcare, and one that better fits the perspective of clinical practice.

This hierarchical approach recognizes the greater strength of evidence generated by systematic reviews and multicentre studies because the findings have been derived from multiple populations, settings and circumstances. In all three dimensions of the evaluation of an intervention, these sources provide the most valid and reliable evidence. Importantly, this hierarchy acknowledges that a range of research methods can contribute valid evidence.

The hierarchy provides a guide that helps the determine best evidence; however, factors such as research quality will also exert an influence on the value of the available evidence. Finally, for an intervention to be fully evaluated, evidence on its effectiveness, appropriateness and feasibility will be required.

References

Ball C., Sackett D.L., Phillips B., Haynes B. & Straus S. (1998) *Levels of Evidence and Grading Recommendations*. Centre for Evidence Based Medicine, http://cebm.jr2.ox.ac.uk/index.extras.

Basche C.E. (1987) Focus groups interview: an underutilized research technique for improving theory and practice in health education. *Health Quarterly* **14**, 411–418.

Beaudin C.L. & Pelletier L.R. (1996) Consumer-based research: using focus groups as a method for evaluating quality of care. *Journal of Nursing Care Quality* **10**, 28–33.

Benson K. & Hartz A.J. (2000) A comparison of observational studies and randomised controlled trials. *New England Journal of Medicine* **342**, 1878–1886.

Black N. (1996) Why we need observational studies to evaluate the effectiveness of healthcare. *British Medical Journal* **312,** 1215–1218.

Brewin C.R. & Bradley C. (1989) Patient preferences and randomised clinical trials. *British Medical Journal* **299,** 313–315.

Canadian Task Force on the Periodic Health Examination. (1979) The periodic health examination. *Canadian Medical Association Journal* **121,** 1193–1254.

Carruthers S.G., Larochelle P., Haynes R.B., Petrasovits A. & Schiffrin E.L. (1993) Report of the Canadian Hypertension Society consensus conference: 1. Introduction. *Canadian Medical Association Journal* **149,** 289–293.

Chalmers T.C., Celano P., Sacks H.S. & Smith H. (1983) Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine* **309,** 1358–1361.

Colditz G.A., Miller J.N. & Mosteller F. (1989) How study design affects outcomes in comparisons of therapy. I: Medical. *Statistics in Medicine* **8,** 441–454.

Concato J., Shah N. & Horwitz R.I. (2000) Randomised controlled trials, observational studies and the hierarchy of research designs. *New England Journal of Medicine* **342,** 1887–1892.

Cook D.J., Guyatt G.H., Laupacis A. & Sackett D.L. (1992) Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* **102,** 305s–311s.

Cook D.J., Guyatt G.H., Laupacis A., Sackett D.L. & Goldberg R.J. (1995) Clinical recommendations using levels of evidence for antithrombotic agents. *Chest* **108,** 227s–230s.

Cook D.J., Mulrow C.D. & Haynes B. (1998) Synthesis of best evidence for clinical decisions. In: *Systematic Reviews: Synthesis of Best Evidence for Health Care Decisions* (eds Mulrow C.D. & Cook D.). American College of Physicians, Philadelphia.

Dawson-Saunders B. & Trapp R.G. (1994) *Basic and Clinical Biostatistics*. Prentice Hall International, London.

Elwood M. (1998) *Critical Appraisal of Epidemiological Studies and Clinical Trials*, 2nd edn. Oxford University Press, Oxford.

Feinstein A.R. (1989) Epidemiologic analysis of causation: the unlearned scientific lessons of randomised trials. *Journal of Clinical Epidemiology* **42,** 481–489.

Guyatt G.H., Sackett D.L., Sinclair J.C., Hayward R., Cook D.J. & Cook R.J. (1995) Users guide to the medical literature: IX. A method for grading healthcare recommendations. *JAMA* **274,** 1800–1804.

Guyatt G.H., Haynes R.B., Jaeschke R.Z., Cook D.J., Green L., Naylor C.D., Wilson M.C. & Richardson W.S. (2000) Users guide to the medical literature XXV. Evidence-based medicine: Principles for applying the users guides to patient care. *JAMA* **284,** 1290–1296.

Horwitz R.I., Viscoli C.M., Clemens J.D. & Sadock R.T. (1990) Developing improved observational methods for evaluating therapeutic effectiveness. *American Journal of Medicine* **89,** 630–638.

McKee M., Britton A., Black N., McPherson K., Sanderson C. & Bain C. (1999) Interpreting the evidence: choosing between randomised and non-randomised studies. *British Medical Journal* **319,** 312–315.

Meltzer S., Leiter L., Daneman D., Gerstein H.C., Lau D., Ludwig S., Yale J., Zinman B. & Lillie D. (1998) 1998 clinical practice guidlines for the management of diabetes in Canada. *Canadian Medical Association Journal* **159,** S1–S29.

Meyer J. (2000) Using qualitative methods in health related action research. *British Medical Journal* **320,** 178–181.

Miller J.N., Colditz G.A. & Mosteller F. (1989) How study design affects outcomes in comparisons of therapy. II. Surgical. *Statistics in Medicine* **8,** 455–466.

Muir Gray J.A. (1997) *Evidence-Based Healthcare*. Churchill Livingstone, New York.

Mulrow C.D. (1987) The medical review article: state of the science. *Annals of Internal Medicine* **106,** 485–488.

Mulrow C.D. & Oxman A.D. (1997) *Cochrane Collaboration Handbook* (database on disk and CDROM). The Cochrane Library, The Cochrane Collaboration, Oxford, Updated Software.

NHMRC (1995) *Guidelines for the Development and Implementation of Clinical Guidelines*, 1st edn. Australian Government Publishing Service, Canberra.

NHS Centre for Reviews and Dissemination (1996) *Undertaking Systematic Reviews of Research on Effectiveness. CRD Guidelines for Those Carrying Out or Commissioning Reviews*. University of York, York.

Robinson N. (1999) The use of focus group methodology: with selected examples from sexual health research. *Journal of Advanced Nursing* **29,** 905–913.

Sackett D.L. (1986) Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* **89,** 2s–3s.

Sackett D.L., Richardson W.S., Rosenberg W. & Haynes R.B. (1997) *Evidence Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, New York.

Van der Zalm J.E. (2000) Hermeneutic-phenomenology: providing living knowledge for nursing practice. *Journal of Advanced Nursing* **31,** 211–218.

Van Manen M. (1990) *Researching Lived Experience: Human Science for an Action Sensitive Pedagogy*. State University of New York, London.

Wilson M.C., Hayward R.S.A., Tunis S.R., Bass E.B. & Guyatt G. (1995) Users guide to the medical literature. VIII. How to use clinical practice guidelines; B. What are the recommendations and will they help you in caring for your patients. *JAMA* **274,** 1630–1632.

Woolf S.H., Battista R.N., Anderson G.M., Logan A.G. & Wang E. (1990) Assessing the clinical effectiveness of preventative maneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *Journal of Clinical Epidemiology* **43,** 891–905.