Public Health Nutrition: 7(2), 279–284 DOI: 10.1079/PHN2003535

History of the concept of 'levels of evidence' and their current status in relation to primary prevention through lifestyle interventions

A Kroke[1,2,3,*], H Boeing[2], K Rossnagel[1] and SN Willich[1]
[1]Institute of Social Medicine, Epidemiology, and Health Economics, Charité, Berlin, Germany:
[2]Department of Epidemiology, German Institute of Human Nutrition, Potsdam–Rehbrücke, Germany:
[3]Research Institute of Child Nutrition, Heinstück 11, D-44225 Dortmund, Germany

Submitted 15 July 2002: Accepted 1 April 2003

Abstract

Primary prevention is a major option to reduce the burden of chronic disease in populations. Because lifestyle interventions have proved to be effective, lifestyle recommendations including nutritional advice are made abundantly. However, both their credibility and their effectiveness are often considered not to be high. Therefore, scientific evidence should form the basis of recommendations and, as in clinical medicine, a rational approach should be followed for the evaluation of evidence. In this paper, the development and current concepts of 'levels of evidence' as they are applied in clinical medicine are outlined and their impact on evidence-based recommendations is discussed. Next, the question is raised as to how far the existing schemes are applicable to the evaluation of issues pertaining to primary prevention through lifestyle changes. Current schemes were developed mainly for clinical research questions and therefore place major emphasis on randomised controlled trials as the main and most convincing evidence in the evaluation process. These types of study are rarely available for lifestyle-related factors and might even not be feasible to obtain. Arguments are advanced to support the notion that a modification of currently existing 'levels of evidence' as developed for clinical research questions might be necessary. Thereby, one might be able to accommodate the specific aspects of evidence-related issues of recommendations for primary prevention through lifestyle changes, like dietary changes.

Keywords
Public health nutrition
Evidence
Epidemiology

Although not well recognised by the medical system, primary prevention is probably the only long-term option for realistically reducing the disease burden in populations. Primary prevention through changes in lifestyle and behaviour has been proved to be highly effective. Examples are condom use and AIDS, smoking and lung cancer, fruits and vegetables and cancer and cardiovascular disease. In the past, recommendations on prevention through lifestyle and behaviour often were considered as statements of more or less unknown credibility and were usually undefined in terms of quantity. However, as in many other areas of medical research, a rational approach is needed in primary prevention through lifestyle interventions.

An important conceptual issue in this context concerns the type of scientific evidence that is regarded as necessary to give recommendations for primary prevention. This issue has been neglected for many decades. It was considered to be sufficient to claim that even if a recommendation does not prove to be effective in terms of preventing disease occurrence, at least it should not harm. This viewpoint is not acceptable and some scientists claim that recommendations for primary prevention should follow a rigid evaluation process similar to that applied for any therapeutic measure. However, it is obvious that, besides scientific dignity, the criteria to evaluate recommendations might be different from those used for clinical guidelines.

In this context we have investigated the history of the concept of 'levels of evidence', which has become an important aspect of the evaluation procedure for evidence-based recommendations, and discuss the current status in view of its future.

The history of 'levels of evidence'

The inaccessible amount of data published in medical journals has created a strong need to summarise findings in clinical medicine and to come to a conclusion based on the best available empirical evidence. In developing

*Corresponding author: Email Kroke@fke-do.de

recommendations for clinical preventive services, the Canadian Task Force on the Periodic Health Examination applied a hierarchy of evidence to rank their recommendations according to the available type and amount of evidence[1] (Table 1). By using that scheme to derive a ranking of recommendations, this approach appears to have been the first practical application of such levels of evidence.

This scheme was used further to derive a grading of recommendations. The highest grade of recommendation (grade A) was established if level I evidence was available, suggesting that this level should be desired for every recommendation. The authors themselves, however, were surprised how few of the recommendations they were working on could be based on this criterion[2].

Since then, levels of evidence have been used widely in evidence-based medicine. In this context, hierarchies of evidence have been further developed and modified. During the past few years, several organisations have created their own version of a hierarchy of evidence (see Appendix). While in all these hierarchies the lowest level of evidence is given to expert opinion and the highest level of evidence to systematic reviews or meta-analyses of randomised controlled trials (RCTs), there is considerable variation among the categories in between. Common to all of these modifications is the emphasis on RCTs and meta-analyses thereof.

The strong focus on RCTs has resulted from the main application of the systematic literature reviews: the formulation of clinical guidelines. In clinical medicine, the evaluation of treatment effectiveness is best done with placebo-controlled, double-blind trials with randomisation to treatment groups. Medical societies and federal agencies worked for a long time to ensure that this type of study became the 'gold standard' in clinical and pharmaceutical research. In particular, the evidence-based medicine movement (e.g. the Cochrane Collaboration) has raised awareness that, for rational decision-making in clinical practice, systematic reviews are a tool to establish consistency of effects as well as a way to reduce bias and chance effects[3]. Thus, it is not surprising that these types of study have become the basis for high-level evidence-based guidelines in many of the evaluation schemes.

Table 1 The Canadian Task Force on the Periodic Health Examination's hierarchy of evidence[1]

Level of evidence	Type of study
I	At least one RCT
II.1	Well-designed controlled trials without randomisation
II.2	Well-designed cohort or case–control analytical studies, preferably from more than one centre or group
II.3	Multiple time-series studies with or without intervention
III	Opinions of respected authorities, clinical experience, descriptive studies or opinions of expert committees

RCT – randomised controlled trial.

Nowadays, the obvious success in improving the quality of clinical practice by systematic reviews and applying quality criteria has motivated medical societies to apply this scheme developed for clinical practice to all research questions in the medical field, including public health. This (in the clinical context) well-founded grading system based on RCTs is now commonly regarded the one and only way to provide reliable answers to all medical questions. Even though it is stated in Cochrane Collaboration handbooks that reviews of other types of evidence can be helpful for decision-making, especially in areas where RCTs are either not available or not feasible[3], the stigma that everything else beyond RCTs is second- or even third-class evidence and therefore basically not credible is inherent to this not foreseen expansion.

A critical appraisal of the hierarchies of evidence and their application appears necessary, however, because a specific type of research question – mainly the evaluation of therapeutic effects – has driven the development of these hierarchies. This has led to the specific order and inclusion of certain study types. Only recently, levels of evidence have been published which take into account that different medical areas require different sets of levels of evidence. The Canadian Task Force on the Periodic Health Examination differentiated the following research categories and now presents separate hierarchies of evidence for each of these categories[4]:

- Therapy/Prevention/Aetiology/Harm;
- Prognosis;
- Diagnosis; and
- Economic analysis.

Organisations like the Oxford Centre for Evidence-based Medicine have adapted this differentiated system[5]. In addition, the two leading institutions dealing with preventive medicine and related issues – the Canadian Task Force on the Periodic Health Examination and the US Preventive Services Task Force – do not directly link the rating of the quality of evidence to the strength of recommendations they make. Level I evidence does not imply a type A (e.g. highest level) recommendation, nor does a type A recommendation require level I evidence. The US Preventive Services Task Force (see Appendix) rather considers further criteria, such as the burden of suffering from the target condition, the characteristics of the intervention and the effectiveness of the intervention as demonstrated in published research[6].

Discussion

The aim of this brief review on the development of levels of evidence was to demonstrate (1) that these hierarchies are not rigid, and (2) that modifications of these hierarchies

https://doi.org/10.1079/PHN2003535 Published online by Cambridge University Press

and grading systems according to different research questions have been made and still have to be made. The question we should like to discuss here is whether existing schemes for levels of evidence and grading of recommendations are appropriate for recommendations/guidelines dealing with behavioural lifestyle modifications. The reason for raising this question is that epidemiological studies are the main scientific basis for research into lifestyle behaviour as a risk factor for disease. And the vast majority of studies in this area are observational.

In many of the grading schemes presented previously, observational research has been shifted to lower levels of evidence and/or the grading of recommendations attributed only second- or third-level grades to recommendations based on results from observational research. In addition, different types of observational study were often listed together in one group without differentiation of study designs, and often were not presented in their completeness. For example, several hierarchies of evidence do not even mention cohort studies at all[7,8].

The question to be asked, therefore, is whether the downgrading of evidence from observational studies compared with RCTs is justified in all cases.

At this point it seems useful to recall the main underlying rationale for the current hierarchies of evidence, which is then often transported into the grades of recommendations. The hierarchies are based on the ranking of studies according to their susceptibility to bias. That means the hierarchy refers to the internal validity of study designs. Clearly, susceptibility to bias is an important indicator of the internal validity of a study but lack of bias does not inform one about the appropriateness of the study design, its external validity or its relevance to the research question at hand. The current concept of hierarchies implies a trade-off between vulnerability to bias and external validity[9], which favours the internal validity aspect. Therefore, evaluation schemes that are based on issues of study design have been found to be inadequate[10].

If study design *per se* is not a criterion for the ranking and grading, what could be the reason to assume that the downgrading of observational studies might not be appropriate in this context? What is the difference between studies on therapeutic effectiveness or clinical preventive measures, on the one hand, and observational studies on the relationship between lifestyle behaviour and disease risk, on the other?

A therapeutic drug has to be administered by the investigator for its effect to be evaluated; a clinical preventive measure such as a vaccine has to be applied first before its effectiveness can be assessed. As the investigator has to introduce the intervention and to allocate the treatment, the study design has to be experimental.

In contrast, lifestyle interventions often refer to 'common' behaviour in a population. Preventive recommendations usually advise people to enhance a certain behaviour (such as to eat more fruits and vegetables) within commonly occurring ranges of behaviour; or to abstain from a certain behaviour, like to quit or refrain from smoking. Therefore, one has to ask how far experimental study designs are necessary to test the effects of such behaviour if this is already observable in the population. Nobody would demand to scrutinise the recommendation to stop smoking for lung cancer prevention by conducting an RCT. Pure observational research has provided sufficient knowledge to initiate preventive measures and the prevention measures have proved to be successful.

Just as a side remark in this context, it is noteworthy that once a drug or vaccine had been used in a population for some time, observational studies that assessed the effectiveness of that intervention were in most reported cases able to demonstrate basically the same effect estimate as the intervention studies[11,12].

The fact that lifestyle behaviours are observable in populations suggests the assessment of their effect in observational studies. In addition, the above-mentioned example of smoking and lung cancer risk also demonstrates the barriers one would encounter if an experimental study were required to prove the insights from observation. First of all, ethical considerations would not allow one to expose people deliberately to the effects of smoking; and second, such a study would take years if not decades to complete. Other lifestyle recommendations require complex behavioural changes so that cause and effect cannot be clearly linked. A dietary intervention study where the intake of fruits and vegetables is doubled compared with baseline intake levels will have various effects on the composition of the whole diet. Are the observed effects of the intervention due to the increase in fruit and vegetable intake or due to the decrease of other food items that goes along with the intended changes? This might be hard to disentangle. Furthermore, blinding of neither the study participants nor the investigators is possible. In the absence of good exposure markers adherence to the protocol might be difficult to assess. Therefore, even if such studies are performed, the question remains of whether possible biases in observational studies on lifestyle factors are judged to be worse than those that might arise from intervention studies in this field.

Another barrier is the long latency period of many chronic diseases. As hard endpoints such as an incident cancer might take too long to occur, the compelling idea of using precursor lesions of intermediate disease risk markers as the outcome measure in human experimental studies has been brought up. This would require solid data about the relationship of that precursor or marker to the final endpoint and a clear idea about the relevant time of exposure. Only if the putative risk factor is causally linked to the development of the precursor and if this precursor always develops further to the endpoint would a trial be

Table 2 Level of evidence for guideline recommendations of the US Agency for Health Care Policy and Research (AHCPR)[9]

Level of evidence	Type of study	Grade of recommendation
1	Supportive evidence from well-conducted RCTs that include 100 patients or more	A
2	Supportive evidence from well-conducted RCTs that include fewer than 100 patients	A
3	Supportive evidence from well-conducted cohort studies	A
4	Supportive evidence from well-conducted case–control studies	B
5	Supportive evidence from poorly controlled or uncontrolled studies	B
6	Conflicting evidence with the weight of evidence supporting the recommendation	B
7	Expert opinion	C

RCT – randomised controlled trial.

able to give the answers sought for. Similarly, studies on high-risk subjects might, although feasible, not give answers about effects in non-high risk subjects.

In essence, RCTs are in certain instances not feasible or not able to provide the adequate answer. Consequently, the grading of evidence or the grading of recommendations for lifestyle interventions respectively may not allow the application of schemes that are designed for clinical therapeutic or individual preventive measures and that rely solely on RCTs. In these instances it might be prudent to assign equal value to observational studies and RCTs, as done by the US Agency for Health Care Policy and Research (AHCPR)[9]. Their hierarchy of evidence is based on seven levels and three grades of recommendation (A–C) are derived from this (Table 2). Grade A recommendations are based on either RCTs or well-conducted cohort studies. The application of this type of grading also takes into account the different vulnerability of observational studies to biases. Cohort studies are usually considered the superior study design compared with case–control studies, mainly because information and selection biases are less likely to occur.

One approach to cope with the remaining uncertainty in evaluating observational study results on exposure–disease relationships is the application of causal criteria. In contrast to RCTs, which evaluate the effectiveness of a therapeutic measure, observational epidemiological studies seek to clarify causal relationships. Hill proposed causal criteria that include aspects such as consistency, temporality, biological gradient, biological plausibility and coherence[13]. Although these criteria have frequently been criticised[14], they have remained as an aid in drawing inferences from observational epidemiological research. These and other approaches to help interpret observed associations underline the need for scrutiny in the evaluation of the evidence that should be applied. However, neither undetected bias nor confounding nor chance can be entirely ruled out as an explanation for a study finding. Therefore, state-of-the-art approaches for the evaluation of evidence are based on systematic summaries of all available data on a given topic rather than on single studies or results from single study types. For the summary of data, meta-analysis is the approach currently most favoured for summarising study results of both RCTs as well as case–control and cohort studies. Pooled analysis, where the original data rather than the results are combined, is another statistical approach to combine results from different studies and to estimate a summary effect. Currently, the methodological debates continue about how to optimise statistical procedures and a need to adapt existing models to the particularities of observational studies has been described[15].

In conclusion, different areas of research in the medical field have different sets of research questions to be answered and therefore have different types of study design available to investigate these issues. The attempt made here to upgrade the value of certain types of observational research is not meant to discredit the conceptual advantages of experimental study designs. However, RCTs are not feasible or available in all situations, but still an answer is needed and recommendations are required[16]. The widespread notion that only RCTs are a valid basis for type A recommendations might delay or even stop decision-makers in the public health sector from devoting attention or resources to primary preventive measures just because, according to certain schemes, no 'grade A' evidence is available. If the sum of all evidence points in one direction and plausible alternative explanations are not present, the mere fact that 'only' observational studies are available should not automatically preclude one from deriving recommendations. We therefore suggest application of a correspondingly revised hierarchy of evidence and grades of recommendations to be derived therefrom. This means that the evaluation of therapeutic and preventive measures should follow separate schemes; consequently, the suggested grouping of topics as presented by the Canadian Task Force on the Periodic Health Examination (see above) should be refined further.

References

1 Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Canadian Medical Association Journal* 1979; **121**: 1193–254.
2 Goldbloom RB. Weighing the evidence: the Canadian experience. *American Journal of Clinical Nutrition* 1997; **65**(Suppl. 2): 584S–6S.
3 Clark M, Oxman AD, eds. Cochrane Reviewers' Handbook 4.1.0. In: *The Cochrane Library*, Issue 2, 2003.
4 Canadian Task Force on the Periodic Health Examination. http://ctfphc.org. Accessed 9 December 1999.

https://doi.org/10.1079/PHN2003535 Published online by Cambridge University Press

5 Oxford Centre for Evidence-based Medicine. *Levels of Evidence and Grades of Recommendations*. Available at http://www.cebm.jr2.ox.ac.uk/docs/level.html. Accessed 31 August 2001.
6 US Preventive Services Task Force. *Guide to Clinical Preventive Services*, 2nd ed. Washington, DC: US Department of Health and Human Services, 1996.
7 Ärztliche Zentralstelle für Qualitätssicherung (ÄZQ)/Arbeitsgemeinschaft wissenschaftlichen medizihischen Fachgesechaften (AWMF). *Das Leitlinien-Manval*. Köln, ÄZQ/AWMF, 2000.
8 Eccles M, Freemantle N, Mason J. North of England evidence based guideline development project: guideline on the use of aspirin as secondary prophylaxis for vascular disease in primary care. North of England Aspirin Guideline Development Group. *British Medical Journal* 1998; **316**: 1303–9.
9 Hadorn DC, Baker D, Hodges JS, Hicks N. Rating the quality of evidence for clinical practice guidelines. *Journal of Clinical Epidemiology* 1996; **49**: 749–54.
10 Lohr KN, Carey TS. Assessing 'best evidence': issues in grading the quality of studies for systematic reviews. *Joint Commission Journal on Quality Improvement* 1999; **25**: 470–9.
11 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *American Journal of Ophthalmology* 2000; **30**: 688.
12 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine* 2000; **342**: 1887–92.
13 Hill A. The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* 1965; **58**: 295–300.
14 Rothman KJ, Greenland S. *Modern Epidemiology*. Philadelphia, PA: Lippincott-Raven, 1998.
15 Dickersin K. Systematic reviews in epidemiology: why are we so far behind? *International Journal of Epidemiology* 2002; **31**: 6–12.
16 Black N. Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal* 1996; **312**: 1215–18.
17 Scottish Intercollegiate Guidelines Network (SIGN). *SIGN 50: A Guideline Developer's Handbook. Section 6* [online]. Available at http://www.sign.ac.uk/guidlines/fulltext/50/sections5.html. Accessed 23 November 2001.

Appendix – Examples for hierarchies of evidence and grading schemes for recommendations

Table A1 North of England Evidence Based Guideline Development Project[8]

Level	Type of evidence	Strength of recommendation
Ia	Evidence from meta-analysis of RCTs	A
Ib	Evidence from at least one RCT	A
IIa	Evidence from at least one controlled study without randomisation	B
IIb	Evidence from at least one other type of quasi-experimental study	B
III	Evidence from descriptive studies, such as comparative studies, correlation studies and case–control studies	C
IV	Evidence from expert committee reports or opinions or clinical experience of respected authorities, or both	D

RCT – randomised controlled trial.

Table A2 Revised Scottish Intercollegiate Guidelines Network (SIGN) grading system[17]

Level	Type of evidence	Grade of recommendation
1++	High-quality meta-analyses, systematic reviews of RCTs or RCTs with a very low risk of bias	A*
1+	Well-conducted meta-analyses, systematic reviews of RCTs or RCTs with a very low risk of bias	A*
1−	Meta-analyses, systematic reviews of RCTs or RCTs with a high risk of bias	
2++	High-quality systematic reviews of case–control or cohort studies *or* High-quality case–control or cohort studies with a very low risk of confounding, bias or chance, and a high probability that the relationship is causal	B†
2+	Well-conducted case–control or cohort studies with a low risk of confounding, bias or chance, and a moderate probability that the relationship is causal	C‡
2−	Case–control or cohort studies with a high risk of confounding, bias or chance, and a significant risk that the relationship is not causal	
3	Non-analytical studies, e.g. case reports, case series	D§
4	Expert opinion	D§

RCT – randomised controlled trial.
*Grade A: at least one meta-analysis, systematic review or RCT rated as 1++ and directly applicable to the target population *or* A systematic review of RCTs or a body of evidence consisting principally of studies rated 1+ directly applicable to the target population and demonstrating overall consistency of results.
†Grade B: a body of evidence including studies rated as 2++ directly applicable to the target population and demonstrating overall consistency *or* Extrapolated evidence from studies rated 1++ or 1+.
‡Grade C: a body of evidence including studies rated as 2+ directly applicable to the target population and demonstrating overall consistency of results *or* Extrapolated evidence from studies rated 2++.
§Grade D: evidence level 3 or 4 *or* Extrapolated evidence from studies rated 2+.

https://doi.org/10.1079/PHN2003535 Published online by Cambridge University Press

Table A3 US Preventive Services Task Force rating system[6]

Rating of the quality of evidence	
I	Evidence obtained from at least one properly randomised controlled trial
II.1	Evidence obtained from well-designed controlled trials without randomisation
II.2	Evidence obtained from well-designed cohort or case–control analytical studies, preferably from more than one centre or research group
II.3	Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence
III	Opinions of respected authorities, based on descriptive studies of clinical experience and case reports or the reports of expert committees
Grading of the strength of recommendations*	
A	There is good evidence to support the recommendation that the condition be specifically considered in a periodic health examination
B	There is fair evidence to support the recommendation that the condition be specifically considered in a periodic health examination
C	There is insufficient evidence to recommend for or against the inclusion of the condition in a periodic health examination, but recommendations may be made on other grounds
D	There is fair evidence to support the recommendation that the condition be excluded from consideration in a periodic health examination
E	There is good evidence to support the recommendation that the condition be excluded from consideration in a periodic health examination

* Determination of the quality of evidence (i.e. 'good', 'fair' and 'insufficient') in the *strength of recommendations* was based on a systematic consideration of three criteria: the burden of suffering from the target condition, the characteristics of the intervention and the effectiveness of the intervention as demonstrated in published clinical research. Effectiveness of the intervention received special emphasis. In reviewing clinical studies, the Task Force used strict criteria for selecting admissible evidence and placed emphasis on the quality of study designs. In rating the *quality of evidence*, the Task Force gave greater weight to those study designs that, for methodological reasons, are less subject to bias and inferential error. The above-mentioned rating system was used.

https://doi.org/10.1079/PHN2003535 Published online by Cambridge University Press