GRADING – levels of evidence

Derek Richards
Editor

As this journal changes, it is worth highlighting one the key elements of the Summaries we publish in Evidence-based Dentistry, namely the assignment of levels of evidence.

We assign levels of evidence to each main Summary published in *Evidence-based Dentistry*, with the exception of guidelines which contain a mix of levels and therefore present more of a challenge. The system we use in the journal is based on that employed by the Oxford Centre for Evidence-based Medicine (OCEBM) as shown in Table 1.

The level of evidence we assign is highlighted using our evidence graphic (Figure 1). We will continue to use this system for the present, but it is worth mentioning some of the work that has taken place in the area over the past few years that may change the way we assign levels of evidence in the journal.

One of the first attempts to explicitly characterise a hierarchy of evidence was made by the Canadian Task Force on the Periodic Health Examination in 1979,[1] to link their healthcare recommendations with the strength of underlying evidence. Since then, Holger *et al.*[2] have identified more than 100 other groups that have used various systems of codes to communicate grades of evidence and recommendations. Glasziou and colleagues[3] subsequently identified five issues that they believed should be addressed when looking at alternative approaches to identifying reliable evidence:

- **Different types of question require different types of evidence.** For example, randomised controlled trials can give good estimates of treatment effects but poor estimates of prognosis.
- **Systematic reviews are preferable:** studies, with rare exceptions, should not be interpreted in isolation, so pooling of study findings using standardised reporting is preferable.
- **Level alone should not be used to grade evidence.** Although this approach helps to justify study selection, a number of disadvantages were identified by these authors, eg, levels may mean different things to different readers, and novel or hybrid approaches are not easily accommodated. This can lead to anomalous rankings, where a systematic review (usually the highest level) that is based on a few small poor quality trials might be placed above a large, well-conducted, multicentre trial.
- **What if there are no systematic reviews?** Systematic reviews are only available for a small number of topics so whatever evidence is found should be clearly described.
- **Balanced assessment should draw on a variety of research.** Even if the effectiveness of any particular treatment has good systematic evidence, data about potential harm is likely to come from cohort or case–control studies: risk–benefit assessments thus need to draw on a variety of research types.

Evidence Graphic	Evidence Level	Therapy/Prevention/ Aetiology/Harm
3A **2C** **2B** **2A** **1B** **1A**	1A	SR (with homogeneity*) of RCTs
3A **2C** **2B** **2A** **1B** 1A	1B	Individual RCT (with narrow Confidence Interval)
3A **2C** **2B** **2A** 1B 1A	2A	SR (with homogeneity*) of cohort studies
3A **2C** **2B** 2A 1B 1A	2B	Individual cohort study (including low quality RCT; e.g. <80% follow-up)
3A **2C** 2B 2A 1B 1A	2C	Ecological studies
3A 2C 2B 2A 1B 1A	3A	SR (with homogeneity*) of case-control studies

* By homogeneity we mean a systematic review that is free of worrisome variations (heterogeneity) in the directions and degrees of results between individual studies. Not all systematic reviews with statistically significant heterogeneity need be worrisome, and not all worrisome heterogeneity need be statistically significant.

Figure 1. The Evidence-based Dentistry evidence graphic.

These authors suggested that there were two broad options to address these concerns; to extend and improve existing hierarchies, or to abolish evidence hierarchies and levels of evidence and concentrate instead on teaching practitioners general principles of research so that they can use these principles to appraise the quality and relevance of particular studies. I would suggest that both are necessary.

In 2004, the GRADE (Grading of Recommendations Assessment, Development and Evaluation) working group published a critical appraisal of the six most prominent systems for grading levels of evidence and

Table 1. Simplified version of the Oxford Centre for Evidence-based Medicine levels of evidence table*

	Type of Question				
Evidence level	**Therapy/ prevention, aetiology/ harm**	**Prognosis**	**Diagnosis**	**Differential diagnosis/ symptom prevalence study**	**Economic and decision analyses**
1a	Systematic review of RCT	Systematic review of cohort studies	Systematic review of level 1 diagnostic studies	Systematic review of prospective cohort studies	Systematic review of level 1 economic studies
1b	Individual RCT with narrow confidence intervals	Individual inception cohort study with ≥80% followup	Validating cohort study with good reference standards	Prospective cohort study with good followup	Analysis based on clinically sensible costs or alternatives; systematic review of the evidence. Multiway sensitivity analysis included
2a	Systematic review of cohort studies	Systematic review of either retrospective cohort studies or untreated control groups in RCT	Systematic review of level >2 diagnostic studies	Systematic review of level ≥2b studies	Systematic review of level >2 economic studies
2b	Individual cohort study (including low quality RCT; eg, <80% followup)	Retrospective cohort study or followup of untreated control patients in RCT	Exploratory cohort study with good reference standards	Retrospective cohort study or poor followup	Analysis based on clinically sensible costs or alternatives; limited review of the evidence, or single studies. Multiway sensitivity analysis included
2c	'Outcomes' research; ecological studies	'Outcomes' research		Ecological studies	Audit or 'outcomes' research
3a	Systematic review of case–control studies		Systematic review of level ≥3b studies	Systematic review of level ≥3b studies	Systematic review of ≥3b studies

RCT, Randomised controlled trial.
*Full version available from Oxford Centre for Evidence-based Medicine website (www.cebm.net/levels_of_evidence.asp)

strength of recommendations,[4] as follows:

- The American College of Chest Physicians
- Australian National Health and Medical Research Council
- OCEBM
- Scottish Intercollegiate Guidelines Network
- US Preventive Services Task Force
- US Task Force on Community Preventive Services

Table 2. GRADE: quality of evidence and definitions

Quality of evidence	Definitions
High quality	Further research is very unlikely to change our confidence in the estimate of effect
Moderate quality	Further research is likely to have an impact on our confidence in the estimate of effect and may change the estimate
Low quality	Further research is very likely to have an impact on our confidence in the estimate of effect and is likely to change the estimate
Very low quality	Any estimate of effect is very uncertain

The working group found that there was poor agreement about the sense of the systems; all of the systems used were considered to have important shortcomings when attempting to grade levels of evidence and the strength of clinical recommendations. There was agreement that the OCEBM system worked well for all four types of questions (effectiveness, harm, diagnosis and prognosis) considered for the appraisal, although it was not without its faults.

This critical appraisal examined both the way these six systems rank the evidence and how they then grade the strength of clinical recommendations. A number of key conclusions were drawn, and a new scheme proposed. This has been adopted by the GRADE group to develop a new rating of quality and strength of evidence (Table 2).[5,6]

The GRADE approach to linking evidence and clinical recommendations has much to recommend it and it is likely that this will be an important system in the future — particularly in guideline development. There are of course differences between the role of this journal and guideline development: *Evidence-based Dentistry* identifies good quality articles and provides a commentary from a practitioner working in the area, whereas guidelines (particularly the better ones) are developed by a group that includes a number of topic specific and methodology experts. Guidelines groups are likely to have access to a very wide knowledge base and are thus well placed to apply the GRADE definitions effectively; more so than the smaller number of people employed in developing and preparing summaries for this journal. Consequently, we will continue to rate studies individually using the OCEBM approach (Table 1) for the foreseeable future. Readers who would like more information on GRADE can find this on their website (www.gradeworkinggroup.org).

1. Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Can Med Assoc J* 1979; **121:** 1193–1254.
2. Guyatt GH, Oxman AD, Vist GE, *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Br Med J* 2008; **336:** 924–926.
3. Glasziou P, Vandenbroucke JP, Chalmers I. Assessing the quality of research. *Br Med J* 2004; **328:** 39–41.
4. Atkins D, Eccles M, Flottorp S, *et al.* Systems for grading the quality of evidence and the strength of recommendations. I. Critical appraisal of existing approaches. The GRADE Working Group. *BMC Health Services Res* 2004; **4:** 38.
5. Atkins D, Briss PA, Eccles M, *et al.* Systems for grading the quality of evidence and the strength of recommendations. II. Pilot study of a new system. *BMC Health Services Res* 2005; **5:** 25.
6. Guyatt GH, Oxman AD, Vist GE, *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Br Med J* 2008; **336;** 924–926.

Evidence-Based Dentistry (2009) **10**, 24–25.
doi:10.1038/sj.ebd.640636