Not all evidence is created equal — so what is good evidence?

Derek Richards
Director, Centre for Evidence-based Dentistry, Oxford, UK

We are living in the information age, bombarded from every side by bits and bytes of information. How do we know how good any of it is? One of the aims of *Evidence-Based Dentistry* is to help the practitioner identify the best evidence. Therefore, we identify here the differing levels of evidence and explain how we will be using these in the journal in future.

Evidence-Based Dentistry (2003) **4,** 17–18. doi:10.1038/sj.ebd.6400160

In our practice, we derive information from a wide array of sources ranging from our own experience to high-quality research. All of this can be described as evidence. Indeed, the essence of the evidence-based approach is to use the evidence from all sources in order to provide the best outcome for the patient. Nevertheless, some evidence is better, stronger or more valid than the rest.

An earlier article[1] highlighted three questions you should ask of each paper:

- Is the study valid?
- What are the results?
- Are the results relevant?

In terms of strength of evidence, it is the validity of a piece of evidence that is important. The validity of a study is the extent to which its design and conduct are likely to prevent systematic errors or bias.[2] Therefore, the more valid a piece of evidence, the greater its strength and the more secure you can feel making treatment decisions based on it.

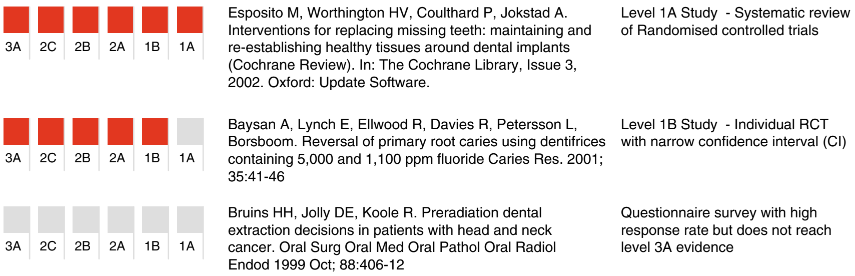
The need to develop a method of ranking the validity of evidence was initially developed by Fletcher and Sackett while working on the Canadian Task Force on Periodic Health Examination.[3] The result was a table of levels of evidence and related "grades of recommendation" for advice based on the levels of evidence. These initial levels and grades have been widely adopted, often in a slightly modified form, by agencies such as the Scottish Intercollegiate Guideline Network (SIGN) and other bodies that develop guidelines and evidence-based publications.

The initial levels and grades were criticised, however, for their therapeutic/preventive orientation. Consequently, the need to develop similar levels for diagnostic, prognostic, harm and economic studies led a group of people associated with the Centre for Evidence-based Medicine to develop a more complete level-of-evidence table (see Table 1), with associated grades of recommendation (see notes to Table 1). The version printed below was last modified in May 2001, but it is also available on the Centre for Evidence-based Medicine website (www.cebm.net) where it is under constant review, so readers should visit the site from time to time to check for changes.

How we use these levels in the *Evidence-Based Dentistry* journal

For *Evidence-Based Dentistry*, we conduct regular searches of the dental and some medical journals to identify possible articles to include in our summary section. The articles we select for inclusion in the journal focus primarily on evidence of level 2a or above although, on occasion, we will include studies down to level 3a evidence as shown in Table 1. Studies below this level are not considered for the journal. The levels of evidence indicated in Table 1 have a narrow focus but, as noted below, the levels may be modified by the addition of plus or minus signs. For example an individual randomised controlled trial with narrow confidence intervals would be rated as a level 1b study. However if the confidence intervals were wide and/or there were other quality questions over the study, this would then be rated level 1b−.

To assist the reader in identifying the level of evidence of a paper, we will in future be including a simple visual device similar to a visual analogue scale. This device will be found at the top of each of the summary papers. An example is shown in Figure 1.



Esposito M, Worthington HV, Coulthard P, Jokstad A. Interventions for replacing missing teeth: maintaining and re-establishing healthy tissues around dental implants (Cochrane Review). In: The Cochrane Library, Issue 3, 2002. Oxford: Update Software.

Level 1A Study - Systematic review of Randomised controlled trials

Baysan A, Lynch E, Ellwood R, Davies R, Petersson L, Borsboom. Reversal of primary root caries using dentifrices containing 5,000 and 1,100 ppm fluoride Caries Res. 2001; 35:41-46

Level 1B Study - Individual RCT with narrow confidence interval (CI)

Bruins HH, Jolly DE, Koole R. Preradiation dental extraction decisions in patients with head and neck cancer. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 1999 Oct; 88:406-12

Questionnaire survey with high response rate but does not reach level 3A evidence

Figure 1. Examples of graphic device to indicate levels of evidence.

1. Richards D. Is it worth reading this paper? EBD 2000; 2:50–52.
2. Moher D, Jadad A, Nichol G, Penman M, Tugwell T, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trial 1995; 16:62–73.
3. Canadian Task Force on the Periodic Health Examination, The periodic health examination. Can Med Assoc J 1979; 121:1193–1254.

Evidence Based Dentistry (2003) **4,** 17–18.
doi:10.1038/sj.ebd.6400160

Table 1. Oxford Centre for Evidence-based Medicine levels of evidence (May 2001).

Level	Therapy/prevention/ aetiology/harm	Prognosis	Diagnosis	Differential diagnosis/ symptom prevalence study	Economic and decision analyses
1a	SR (with homogeneity[a]) of RCTs	SR (with homogeneity[a]) of inception cohort studies; CDR validated in different populations	SR (with homogeneity[a]) of level 1 diagnostic studies; CDR with 1b studies from different clinical centres	SR (with homogeneity[a]) of prospective cohort studies	SR (with homogeneity[a]) of level 1 economic studies
1b	Individual RCT with narrow confidence interval (CI)	Individual inception cohort study with □ 80% follow-up; CDR validated in a single population	Validating[b] cohort study with good reference standards[c]; or CDR tested within one clinical centre	Prospective cohort study with good follow-up[d]	Analysis based on clinically sensible costs or alternatives; SR(s) of evidence; and including muli-way sensitivity analysis
1c	All or none[e]	All or none case-series	Absolute SpPins* and SnNouts**	All or none case-series	Absolute better-value or worse-value analyses[f]
2a	SR (with homogeneity[a]) of cohort studies	SR (with homogeneity[a]) of retrospective cohort studies or untreated control groups in RCTs	SR (with homogeneity[a]) of level >2 diagnostic studies	SR (with homogeneity[a]) of 2b and better studies	SR (with homogeneity[a]) of level >2 economic studies
2b	Individual cohort studies (including low quality RCT, eg <80% follow-up)	Retrospective cohort study or follow-up of untreated control patients in an RCT; derivation of CDR or validated on split-sample[g] only	Exploratory cohort study[h] with good reference standards; CDR after derivation, or validated only on split-sample[g] or databases	Retrospective cohort study, or poor follow-up	Based on clinically sensible costs/alternatives; limited review(s) of the evidence, or single studies; and including multi-way sensitivity analyses
2c	"Outcomes" Research; Ecological studies	"Outcomes" Research		Ecological studies	Audit or "Outcomes" research
3a	SR (with homogeneity[a]) of case-controlled studies		SR (with homogeneity [a]) of 3b and better studies	SR (with homogeneity [a]) of 3b and better studies	SR (with homogeneity [a]) of 3b and better studies
3b	Individual case-controlled studies		Non-consecutive study; or without consistently applied reference standards	Non-consecutive cohort study, or very limited population	Analysis based on limited alternatives or costs, poor-quality estimates of data, but including sensitivity analyses incorporating clinically sensible variations
4	Case-series (and poor-quality cohort and case-controlled studies[i])	Case-series (and poor-quality prognostic cohort studies[j])	Case-controlled study, poor or non-independent reference standard	Case-series or superseded reference standards	Analysis with no sensitivity analyses
5	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"				

Produced by R. Phillips, C. Ball, D. Sackett, D. Badenoch, S. Straus, B. Haynes and M. Dawes since November 1998. Reproduced by permission from Centre for Evidence-based Medicine, Oxford.

Note Users can add a (–) to denote a study in a particular level that fails to provide a conclusive answer because of either: a single result with a wide CI (such that, for example, an absolute risk reduction (ARR) in an RCT is not statistically significant but whose CI fails to exclude clinically important benefit or harm); or an SR with troublesome (and statistically significant) heterogeneity; or the evidence is inconclusive, and therefore can only generate grade d recommendations.

SR, systematic review; RCT, randomised controlled trial; CDR, clinical decision rule algorithms or scoring systems that lead to a prognostic estimation or a diagnostic category.

*SpPins: An 'Absolute SpPin' is a diagnostic finding whose Specificity is so high that a Positive result rules-in the diagnosis.

**SnNouts: An 'Absolute SnNout' is a diagnostic finding whose Sensitivity is so high that a Negative result rules-out the diagnosis.

[a] By homogeneity we mean a systematic review free of worrisome variations (heterogeneity) in the direction and degree 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. As noted above, studies displaying worrisome heterogeneity should be tagged (–) along with their designated level.

[b] Validating studies test the quality of a specific diagnostic test, based on prior evidence. An exploratory study collects information and trawls the data (eg, using a regression analysis) to find which factors are significant.

[c] Good reference standards are independent of the test, and applied blindly or objectively applied to all patients. Poor reference standards are haphazardly applied, but still independent of the test. Use of a nonindependent reference standard (where the test is included in the reference, or where the testing affects the reference) implies a level 4 study.

[d] Good follow-up in a differential diagnosis study is >80%, with adequate time for alternative diagnoses to emerge (eg, 1–6 months acute, 1–5 years chronic).

[e] Met when all patients died before the treatment became available, but some now survive on it or when some patients died before the treatment became available, but none now die on it.

[f] Better-value treatments are clearly as good but cheaper, or better at the same or reduced cost. Worse-value treatments are as good and more expensive, or worse and equally or more expensive.

[g] Split-sample validation is achieved by collecting all the information in a single tranche, then artificially dividing this into "derivation" and "validation" samples.

[h] Validating studies test the quality of a specific diagnostic test, based on prior evidence. An exploratory study collects information and trawls the data (eg, using a regression analysis) to find which factors are significant.

[i] Poor-quality cohort studies are ones that failed to clearly define comparison groups and/or failed to measure exposures and outcomes in the same (preferably blinded), objective way in both exposed and nonexposed individuals and/or failed to identify or appropriately control known confounders and/or failed to carry out a sufficiently long and complete follow-up of patients. Poor-quality case–control studies failed to define clearly comparison groups and/or failed to measure exposures and outcomes in the same (preferably blinded), objective way in both cases and controls and/or failed to identify or appropriately control known confounders.

[j] By poor-quality prognostic cohort study, we mean one in which sampling was biased in favour of patients who already had the target outcome, or the measurement of outcomes was accomplished in <80% of study patients, or outcomes were determined in an unblinded, non-objective way, or there was no correction for confounding factors.