Chapter 2
The Hierarchy of Evidence: From Unsystematic Clinical Observations to Systematic Reviews

Mohamed B. Elamin and Victor M. Montori

Keywords Evidence-based medicine • Hierarchy of evidence • Study design

Any observation in nature is evidence [1]. The human brain is infinite in its ability to draw cause-and-effect inferences from these observations. Unfortunately, these inferences are open to cognitive errors. The scientific method, a method that relies on observations in nature and on evidence, has evolved to minimize error, both random (or due to chance) and systematic error or bias. A key principle of evidence-based medicine is the recognition that not all evidence is similarly protected against error, and that decisions that rely on evidence would be more confident when the evidence is more protected against bias by virtue of the methods used [2]. Thus, a fundamental principle of evidence-based medicine is the recognition of a hierarchy of evidence.

In this chapter, we will review the different approaches the scientific method has evolved to protect evidence from bias. We will then review the evolution of how methodologists have built hierarchies of evidence and note the limitations and merits of these approaches. While this field continues to move forward, we will finish describing what we think represents the state-of-the-art approach to hierarchies of evidence at the time of writing this chapter.

What Is a Hierarchy of Evidence?

To the extent that the evidence is protected against bias it would lead to more confident decision making [2]. Using "risk of bias" as an organizing principle results in a hierarchy of evidence that places studies with better protection against bias at the

V.M. Montori (✉)
Knowledge and Evaluation Research Unit, Department of Medicine,
Division of Endocrinology, Mayo Clinic, Rochester, MN, USA
e-mail: kerunit@mayo.edu

J.G. Burneo et al. (eds.), *Neurology: An Evidence-Based Approach*,
DOI 10.1007/978-0-387-88555-1_2, © Springer Science+Business Media, LLC 2012

top and less-protected evidence at the bottom. "Risk of bias" may not be the only desirable organizing principle of all available hierarchies, but we will focus on it and on the ability to apply evidence to the care of the individual patient when we discuss the position of different forms of evidence on a hierarchy of evidence.

Unsystematic Observations

Imagine that you are seeing a patient diagnosed with multiple sclerosis (MS). One of your clinical preceptors recommended using cyclophosphamide in the treatment of these patients. He had seen many patients improve on this drug and considered the drug both greatly efficacious and quite safe given the patients' dilemma.

Indeed, prior to the advent of evidence-based medicine, unsystematic personal observations from experienced clinicians carried great weight in shaping the practice and teaching of medicine. These observations are the subject of a number of biases introduced by psychological and cognitive processes that make recall and summary of one's experiences suspect. Clinicians interested in exploring these biases can review the work of Kahnemann and Tversky, and of Gigerenzer and colleagues [3–6]. These biases were recognized in research practice prior to clinical practice and the need for methods that will limit the possibility of error, both random and systematic, arose. Indeed, in many hierarchies, unsystematic personal observations often take the lowest or least trustworthy position and are often mistakenly considered "expert opinion." Opinions about observations should not be confused with the observations themselves (evidence), and experts can derive their opinions from any level of the hierarchy of evidence. Thus, expert opinion should not be part of any hierarchy of evidence.

Moving from your memories of what your teacher may have indicated, you seek to look at the body of scientific studies about the risk and benefits of available therapies. As a part of this effort, you decide to search for evidence investigating the use of cyclophosphamide in patients with MS and you find some studies describing the basis by which cyclophosphamide exerts its effect on MS.

Physiology and Mechanistic Studies

Physiology studies, both descriptive and experimental, provide us with the support we need to understand why, for instance, cyclophosphamide and other immunosuppressive regimens might help ameliorate MS symptoms. Searching for physiology studies, you find one of many mechanistic studies published that may potentially help you understand the pathogenesis of MS. This study found increased levels of interleukin-12 in patients with progressive MS compared to controls [7]. How strong is the evidence from physiology studies to support clinical treatment decisions?

There are multiple experiences in which mechanistic explanations have failed to predict outcomes in patients. Trying to answer a question of whether clofibrate in men without clinically evident ischemic heart disease will affect mortality, a before–after physiology study examined the effect of clofibrate on total cholesterol [3]. Patients were given 750–1,500 mg of the drug for 4 weeks after which a significant reduction in total cholesterol level was achieved in 30 out of 35 treated patients. Moreover, the tolerance to the drug was excellent and was not associated with any observed side effects. The positive expectation suggested by these results was shattered by the results of a randomized trial in men randomized to receive either clofibrate or placebo. After a mean follow-up of 9.6 years, the drug increased the risk of death by 25% ($P<0.05$) despite reducing cholesterol levels and the risk of ischemic heart disease (20%, $P<0.05$) [4].

Case Reports and Case Series

Case reports describe individual patients who showed an unusual or an unexpected event, either favorable or unfavorable. These cases may lead to new clinical hypotheses and further clinical care research. Furthermore, case reports and case series are extremely useful in documenting rare events, which may have been obscured in other study designs, as we will discuss later. This makes them of particular importance in studies of harm i.e., unwanted events, which could not be readily studied in an intentional manner, or because of their rarity cannot be studied prospectively.

In your literature search, you stumble over a case report describing a 48-year-old woman, similar to your patient, describing complete remission of MS with a dose of 3,800 mg of cyclophosphamide, which interestingly was given accidentally to that patient [5]. At this point, you were impressed with the results of that case report and how it fits with the biological understanding of both the disease and the medication. Because case reports describe individual patients with no comparison, it is difficult to know that this patient improved *because* of this treatment. This makes case reports highly susceptible to erroneous cause and effect inferences and does not protect against possible confounders, including the passage of time (e.g., spontaneous waxing and waning of disease manifestations over time).

You also find a series of five patients with MS who are not responding to multiple treatments [6]. These patients were given monthly pulse intravenous cyclophosphamide at a dose of 1 g/m^2 unlike the treatment described in the case you read previously. The authors conclude that "aggressive immunosuppressive therapy may be useful in some rapidly deteriorating refractory patients and further controlled studies should be considered in order to full evaluate this type of treatment as a potential therapy in MS." Evidence of large treatment effects in patients who have not responded to therapy and who otherwise have stable or deteriorating disease is often

compelling, with some residual uncertainty associated with patient expectations (placebo effects), natural history of the disease, and other potential explanations. The search to decrease this residual uncertainty must then continue to ensure that your decision making gains in confidence.

Case-Control Study Design

Case-control studies are best used when conditions are rare and investigators are interested in identifying risk factors for their development. These studies enroll a group of patients with a given outcome or condition. The investigators aim to identify risk factors in a retrospective fashion. These studies identify risk factors as those characteristics that are more common among those with the outcome than in those without the outcome (protective factors can be identified with the inverse association). The key determinants of the validity of these study designs rely on the nature of the comparison group (e.g., do these only differ in their outcome status?), and in the quality of the ascertainment of both exposures and outcomes.

In your patient's medical record you find a history of infectious mononucleosis (IM). You look for studies, which might have reported an association between MS and IM. You find a case-control study of 225 patients with MS and 900 controls matched for age and gender [7]. The researchers compared the mean rates of IM infection per patient in different period of times preceding the onset of MS symptoms. They found that a history of IM was significantly associated with the risk of MS compared to the controls (odds ratio 5.5, 95% CI 1.5, 19.7). This information may help you explain why the patient developed MS. Of interest, observational studies of this nature may have explored many different exposures (a key advantage of case-control studies) and only published the statistically significant associations, some of which may have occurred by chance. Other studies that may have found no association with IM, may not have been published, leaving the reader with the impression that IM *causes* MS. Furthermore, the temporary sequence almost always (perhaps with the exception of genetic risk factors) is difficult to establish such that the risk factor may or may not have occurred prior to the development of the disease, a concern that is further worsened by recall bias. Also, difficult to measure or unexpected risk factors or factors associated with those which were measured or assessed may not be accounted for and as a result these studies could mislead.

Cross-sectional studies seeks coexistence of factors at a point in time and – unlike other observational studies that follow individuals over time – cross-sectional studies report what is present or not at a fixed time period, e.g., prevalence of a condition. Indeed, one of the studies you find is a cross-sectional study of the association between MS-related fatigue and treatment. This study was conducted on 320 patients with MS, of which half of them had a complaint of fatigue [8]. After controlling for several factors, the investigators found no significant association between use of immunosuppressive or immunomodulatory drugs and MS-related fatigue. While these studies suggest that cyclophosphamide may not improve fatigue, you

Table 2.1 Criteria that strengthen causal inferences

Analogy
Plausibility
Consistency
Dose–response
Reversibility
Specificity
Strength
Temporality

Adapted from Hill [9]

must remember that cross-sectional studies do not accurately establish causal relationship between exposures and outcomes with multiple explanations for the presence or absence of association. Consider, for instance the fact that one cannot establish the order in which exposure and outcome occurred when sampling at one point in time – can treatment improve in some patients *and* cause fatigue in others? Can those on treatment report fatigue differently than those on different or no treatment? Studies that follow patients over time may better deal with temporal relationships – it is key to establish that an exposure preceded an outcome in order to make causal inferences. Table 2.1 describes criteria that strengthen causal inferences set forth by Bradford Hill [9].

Cohort Study Designs

A cohort study design, in general, enrolls individuals characterized by their exposure status (as oppose to case-control studies which enroll individuals by their outcome status) and follows them for a period with the expectation that some of them will develop the outcomes of interest. These then allow the investigator to measure the incidence or risk of developing the outcome and compare this risk among those exposed and those unexposed. When the investigator plans the study after the participants have started follow-up, then the cohort study is said to be retrospective (the longitudinal follow-up happened in the past); when the investigator sets up the study before the individuals start follow-up, the cohort study is said to be prospective. The nature of the cohort study, prospective or retrospective, does not determine the quality of the cohort study, although prospective cohort studies offer the investigator greater control over the ascertainment of exposure and outcomes and the opportunity to limit the introduction of bias.

Cohort studies can be setup to follow patients for a long time, which makes them suitable for the study of the natural history of disease and for the detection of uncommon harms of treatment or consequences of disease that occur after long exposures (i.e., postmarketing surveillance). Well-conducted cohort studies may occupy top positions in some hierarchies, as we will discuss later.

Among the many limitations of observational studies is that the exposure (i.e., to the treatment vs. no treatment) occurs by choice rather than by chance. This means that when treatment is associated with outcome, it is not only the treatment but also the reasons the patient received the treatment that are associated with the outcome. For instance, women receiving estrogen were found to have lower cardiovascular risk in prospective cohort studies. Importantly, these women were also of higher socioeconomic status, had better access to healthcare, had healthier habits, and took better care of themselves than women who did not receive estrogen therapy. The ability of these observational studies to account for these factors associated with both treatment and outcome (also known as confounders), was limited and only the randomized trials (which assigned exposure by chance rather than by choice) were able to elucidate the lack of cardiovascular protection afforded by estrogen preparations. Many comparisons have shown, however, that observational studies and randomized controlled trials (RCTs) often agree [10]. The trick here is that sometimes they do not and there is no way to know until the randomized trials are performed.

Two examples vividly reflect the importance of the residual uncertainty that exists when inferences drawn from the results of observational studies (with results supported by strong, often post hoc, biological rationale) go unchecked in a randomized trial. Consider an observational study based on secondary analysis of data obtained from a randomized trial data [11], which found that high-dose aspirin (650–1,300 mg daily) given to patients undergoing carotid endarterectomy was associated with 1.8% risk of perioperative stroke and death compared to 6.9% after low-dose aspirin (0–325 mg daily). Later, the randomized trial showed that high-dose aspirin was associated with an 8.4% risk of stroke, myocardial infarction, or death compared to only 6.2% risk in patients receiving low-dose aspirin ($P=0.03$) [12]. Or consider an observational study assessing the effect of extracranial to intracranial bypass surgery on altering the risk of ischemic stroke, a pre–post examination of 110 patients undergoing the bypass was performed. Stroke rate was 4.3% in 70 patients with transient ischemic attacks undergoing the bypass compared with a rate between 13% and 62% in transient ischemic patients who have not undergone surgery and were reported in other published literature. After a 3-year follow-up of all the 110 patients, stroke rate was 5% [13]. The readers would conclude that extracranial to intracranial bypass led to improvement in the symptoms of all patients. In contrast to this conclusion, an RCT of 1,377 patients studying whether bypass surgery benefits patients with symptomatic atherosclerotic disease of the internal carotid artery, found a 14% increase in the relative risk of stroke in patients undergoing surgery over those treated medically [14].

Randomized Trials

In all previous designs, the exposure was not under the control of the investigator and is thus considered observational. This is in contrast with randomized trials in which investigators randomly assign participants to either intervention or control.

Thus, obligatorily these studies are executed prospectively (making it redundant to describe these as "prospective randomized trials"). A well-conducted trial limits any opportunity for patients, clinicians, or investigators to choose to which arm of the trial the participant will be assigned. This feature (randomization) limits bias by not allowing for selecting patients with different prognosis to go to different trial arms. To protect randomization, trials conceal the allocation sequence from participants and investigators, particularly from investigators assessing the eligibility of patients. The most common form to conceal the allocation sequence is central randomization (by computer or phone, at the pharmacy). Enough participants allow chance to also achieve another goal of randomization (in addition to preventing selection bias), which is to create groups with the same prognosis. This allows the investigators to draw causal inferences linking treatment or control to the different prognoses of these arms at the end of the trial. In addition to having two groups with similar prognosis at baseline, blinding of participants, clinicians, and investigators prevents the introduction of cointerventions that would differ between the arms and offer alternative explanations to the findings of these studies. To preserve this balance in prognosis, it is important that these studies follow the intention to treat principle [15]. This principle states that patients should stay in the arm to which they were randomized throughout the study conduct and analyses. Thus, intention-to-treat trials do not have patients unavailable to ascertain their outcomes (loss to follow-up), do not allow unplanned cross over, and seek to have patients receiving as much of their planned exposure for as long as possible. This will provide an unbiased estimate of the treatment effect.

You find a multicenter, placebo-controlled randomized trial studying the effect of cyclophosphamide and other treatments in patients with MS [16]. After at least 12 months of follow-up, the effects of cyclophosphamide given to MS patients did not statistically differ from patients receiving placebo (35% of treatment failures with cyclophosphamide vs. 29% with placebo). You realize, however, that other randomized trials are available and that they have found different results.

Individual-patient randomized trials can only be used to evaluate the effect of treatment in individual patient with stable conditions for which the candidate treatment can exert a temporary and reversible effect. Individual patient randomized trials (also known as *n*-of-1 trials) require the clinician and patient to use a random sequence to determine treatment order. The patient starts the trial with either the intervention or a matching placebo prepared by a third party, a pharmacist for example. The patient and clinician record the effect of the intervention and ensure patients go through a random sequence of exposure to treatment or placebo, typically 3 times [2]. At the end of the trial, both the physician and the patient will have evidence to determine whether the intervention was beneficial or not. An example of such a study design was an *n*-of-1 study conducted with a patient diagnosed with chronic inflammatory demyelinating polyradiculopathy [12]. Although showing initial improvement of symptoms with subsequent remission and relapses, treatment with prednisolone and azathioprine did not stop the slow disease progression. Evaluation of the use of intravenous immunoglobulin (IVIG) was commenced in a blinded placebo-controlled trial with four treatment cycles, consisting of four infusions,

two IVIG (0.4 g/kg) and two albumin infusions as placebo. Each infusion was given once every 3 weeks over a period of 48 weeks. The neurological outcomes of interest were time to walk 10 m, maximum number of squats in 30 s, and maximum range of ankle dorsiflexion, all of which failed to find a clear treatment effect.

Systematic Reviews and Meta-analyses

Evidence-based medicine requires that decisions be made taken into account the body of evidence, not just a single study [2]. Thus, clinicians should be most interested in studies that systematically and thoroughly search for studies that would answer a focused review question. Candidate studies are assessed using explicit eligibility criteria and those selected are subject to evaluation regarding the extent to which they are protected from bias. Investigators then systematically extract data from these studies and summarize it. When these summaries involve statistical pooling, we then say that the systematic review included a meta-analysis. Of note, meta-analyses could also be conducted on an arbitrary collection (i.e., biased selection) of studies; thus the key methodological features is that evidence collections are systematic and that assess the quality of the included studies; meta-analyses do not improve the quality of the studies summarized and will also reflect any biases introduced in the study-selection process. Thus, clinicians should not look for meta-analyses but for systematic reviews (preferably those that conduct a meta-analysis).

Systematic reviews offer evidence that is as good as the best available evidence summarized by the review [2]. For example, for a given research question, high-quality systematic reviews including high-quality trials would yield stronger inferences than systematic reviews of lower quality trials or well-conducted observational studies. Stronger inferences will also be drawn when the studies in the review show consistent answers or when the inconsistency can be explained (often through subgroup analyses). Thus, systematic reviews contribute by improving the applicability of the evidence, and through meta-analyses, by increasing the precision of the estimates of treatment effect. What systematic reviews and meta-analyses do not achieve is the amelioration of any biases present in the studies summarized.

Another key limitation of systematic reviews is that they often rely on published evidence. The published record is subject to bias to the extent that some studies get published later or never and in obscure journals depending on their results, a phenomenon called publication bias. To minimize the possibility of publication bias, the reviewers can search thoroughly and systematically and contact experts in the field. When the studies are published but select the outcomes that received full attention in the manuscript on the basis of their results, a similar phenomenon, reporting bias, takes place. To minimize reporting bias, reviewers contact authors of these studies to verify the data collected and to ask for pertinent data that may have been omitted in the original publication.

You found a systematic review of RCTs studying the efficacy of cyclophosphamide in patients with MS [17]. The systematic review in general was well conducted; the

search was thorough, study selection and data extraction was done in a duplicate manner, and authors of primary studies were contacted for missing data. The trials included in the study, although randomized, had serious study design limitations. Data regarding the Expanded Disability Status Scale score was extracted from eligible trials to measure the course of the disease. Pooled results showed that, compared to placebo or no treatment, intensive immunosuppression with cyclophosphamide in patients with progressive MS did not significantly prevent progression to long-term disability as defined as evolution to a next step in the disability score. After reading this review, you conclude that cyclophosphamide is not your best treatment option and you start looking for different approaches to treat your patient.

RAND Methodology

This technique is used to combine evidence with expert opinion [18, 19]. It starts by summarizing available evidence addressing a specific issue, usually by conducting systematic reviews. Evidence is then synthesized and distributed to a panel of experts in the field under investigation. The panel members rate predefined indicators first, and then arrange for a face-to-face meeting where deliberations are held. After the meeting, the panel members may change their previous ratings based on the discussions they had.

The strengths of the RAND technique is that it allows panelists to rate indicators based on a systematic summary of the evidence, and that they meet together in order for an open discussion to occur. Potential drawbacks may include possible intimidation of some panel members by more influential members. For practical reasons the panel may not be as large as one would prefer and it would not include patients' participation.

This overview of the different study designs should begin to configure a hierarchy of evidence in the reader's mind. Several elements reviewed here could help configure such a hierarchy: protection against biased inferences of cause and effect, consistency and precision of the estimates, and applicability of the findings. We will now review how the notion of a hierarchy of evidence has evolved. To approach this topic, we will focus mostly on hierarchies for prevention and treatment.

History of the Modern Hierarchy of Evidence

Methodologists, early in the evolution of practice guidelines, thought of designing a guide to different levels of evidence in the form of hierarchies that would link the quality of the evidence to the strength of the recommendations in the guidelines. To our knowledge, the first published modern hierarchy of evidence appeared in the year 1979, authored by the Canadian Task Force on the Periodic Health Examination (Table 2.2) [20].

Table 2.2 The Canadian Task Force's hierarchy of evidence

I.	Evidence obtained from at least one properly RCT.
II-1.	Evidence obtained from well designed cohort or case-control analytic studies, preferably from more than one center or research group.
II-2.	Evidence obtained from comparisons between times or places with or without the intervention. Dramatic results in uncontrolled experiments.
III.	Opinions of respected authorities, based on clinical experience, descriptive studies or reports of expert committees.

Table 2.3 U.S. Preventive Services Task Force's hierarchy of evidence

I.	At least one well-conducted RCT
II-1.	Controlled trials without randomization
II-2.	Well-designed cohort or case-control studies, preferably from multiple sites
II-3.	Multiple time-series with or without intervention
III.	Expert opinion

Adapted from Atkins et al. [21]

Table 2.4 The American College of Chest Physicians' hierarchy of evidence

I.	Evidence from randomized trials with a statistically significant effect.
II.	Evidence from randomized trials with a statistically insignificant effect.
III.	Evidence from nonrandomized concurrent comparisons.
IV.	Evidence from nonrandomized comparisons between current patients who received therapy and former patients.
V.	Evidence from case series without a control group.

This first classification of evidence depended mainly on study design (as a surrogate for the quality of evidence) regardless of the quality of the actual studies. It placed RCT, of any quality, at the top of its hierarchy. It also considered opinions from experts as a level of evidence, though, as we discussed earlier, opinions are not forms of evidence. One can see elements of this first hierarchy in some contemporary ones, including the one used until recently by the US Preventive Services Task Force (Table 2.3) [21–23].

Sackett et al. added to study design the precision of the results (in the hypothesis testing framework) in his hierarchy of evidence in 1989 [24].

When the American College of Chest Physicians drafted its antithrombotic guidelines, they correctly excluded expert opinion as a form of evidence and added meta-analyses (rather than meta-analyses within systematic reviews) to their hierarchy (Table 2.4) [24]. Other hierarchies have used this one as a template (see for instance those by Cook et al., Guyatt et al., and Wilson et al.) [25–28].

In 2000, Guyatt et al. introduced a new hierarchy for studies of therapy (Table 2.5) [1]. For the first time, this hierarchy considered *n*-of-1 trials at the top

Table 2.5 Hierarchy of evidence for treatment decisions

I.	Evidence from n-of-1 randomized trial.
II.	Evidence from systematic reviews of randomized trials.
III.	Evidence from single randomized trials.
IV.	Evidence from systematic reviews of observational studies.
V.	Evidence from single observational studies.
VI.	Evidence from physiological studies.
VII.	Evidence from unsystematic observations.

Table 2.6 The American Academy of Neurology's hierarchy of evidence

I.	Randomized, controlled clinical trial with masked or objective outcome assessment in a representative population. Relevant baseline characteristics are presented and substantially equivalent among treatment groups or there is appropriate statistical adjustment for differences. The following are required: (a) concealed allocation, (b) primary outcome(s) clearly defined, (c) exclusion/inclusion criteria clearly defined, and (d) adequate accounting for drop-outs (with at least 80% of enrolled subjects completing the study) and cross-overs with numbers sufficiently low to have minimal potential for bias.
II.	Prospective matched group cohort study in a representative population with masked outcome assessment that meets b–d above OR a RCT in a representative population that lacks one criteria a–d.
III.	All other controlled trials (including well-defined natural history controls or patients serving as own controls) in a representative population, where outcome is independently assessed, or independently derived by objective outcome measurement.
IV.	Studies not meeting Class I, II or III criteria including: consensus, expert opinion or a case report.

Adapted with permission from the American Academy of Neurology [27]

of the hierarchy. In this hierarchy, systematic reviews of randomized trial and of observational studies were included as separate levels of evidence. These modifications indicate that in addition to bias protection and precision, now applicability (enhanced in both *n*-of-1 trials and systematic reviews) became a guiding principle of these hierarchies.

Different organizations adopted or modified preexisting hierarchies [25]. Others, in addition to evidence of treatment studies, added evidence levels for prognostic, diagnostic, and economic analytic studies [26]. The American Academy of Neurology (ANA) adopted its own classification of evidence (Table 2.6) [27]. ANA classified evidence into four classes (I through IV). RCTs were given the highest level of evidence, but with certain requirements. Expert opinions were included at the bottom of the hierarchy, perpetuating the error of the earlier hierarchies. Again, this hierarchy includes elements of bias protection, precision, and applicability.

The state-of-the-art approach to describe the quality of evidence was recently released by the Grades of recommendation, assessment, development, and evaluation (GRADE) working group. The GRADE system and guidelines will be discussed later in Chap. 3.

Practical Applications

The term evidence-based medicine first appeared in 1991 in an article by Gordon Guyatt in the American College of Physicians' Journal Club [28]. Since then, the concept of practicing and teaching evidence-based medicine has exploded. The fundamental principles of evidence-based medicine are (1) the better the quality of the evidence, the more confident the clinical decision making, and (2) evidence alone does not tell us what to do, but decisions should also incorporate patient preferences and values as well as the patient's clinical and personal context. The central role of the hierarchy of evidence in the practice of evidence-based medicine indicates that it is key when considering policies for clinicians: evidence-based policy makers should rely on the highest quality of evidence.

Professional organizations have embarked on the task of developing clinical practice guidelines to provide helpful recommendations to practicing clinicians, to improve quality of care, and to enhance patient outcomes. By producing guidelines, these organizations seek to emphasize their academic credentials and assert their leadership in areas of primary concern. Given the policy and legal implications of guidelines, state-of-the-art guideline developers should follow rigorous and transparent procedures for formulating recommendations for or against a particular diagnostic or therapeutic intervention [2]. Key to their success is the expectation that clinicians will deliver better care for their patients if they follow guideline recommendations. To achieve this goal, the strongest recommendations should result from considering the highest quality of evidence; the higher the level of evidence, the stronger the recommendations. This can be seen in ANA's classification of recommendations based on the level of evidence into four levels (Table 2.7) [27].

Table 2.7 AAN classification of recommendations for therapeutic intervention[a]

A.	Established as effective, ineffective or harmful (or established as useful/predictive or not useful/predictive) for the given condition in the specified population. (Level A rating requires at least two consistent Class I studies.)[b]
B.	Probably effective, ineffective or harmful (or probably useful/predictive or not useful/predictive) for the given condition in the specified population. (Level B rating requires at least one Class I study or at least two consistent Class II studies.)
C.	Possibly effective, ineffective or harmful (or possibly useful/predictive or not useful/predictive) for the given condition in the specified population. (Level C rating requires at least one Class II study or two consistent Class III studies.)
U.	Data inadequate or conflicting; given current knowledge, treatment (test, predictor) is unproven. (Studies not meeting criteria for Class I–Class III.)

Adapted with permission from the American Academy of Neurology [27]

[a]Please refer to Table 2.6 for classification of studies

[b]In exceptional cases, one convincing Class I study may suffice for an "A" recommendation if (1) all criteria are met, (2) the magnitude of effect is large (relative rate improved outcome >5 and the lower limit of the confidence interval is >2)

Conclusion

The recognition of hierarchies of evidence is a key principle of evidence-based medicine. We have discussed here how the need to protect inferences against error has guided the sophistication of the scientific method from unsystematic observations to very large rigorous experiments. We have also reviewed how policy makers have refined the idea of a hierarchy of evidence that initially set forth risk of bias as the sole organizing principle to current strategies that also consider risk of random error (precision), applicability or directness, publication and reporting bias, and consistency in results across studies as additional features of the evidence base to consider. This increased sophistication has set aside reliance of judgments only at the study level moving to making judgments at the "body of evidence" level. Finally, it has also corrected the initial mistake of confusing opinion (of an expert, of a panel, or otherwise) with the observations (evidence) that support such opinions.

It is helpful also to remember that the recognition of hierarchies of evidence is not the only principle of evidence-based medicine. As such, the application of the evidence into clinical decision making and policy making requires consideration of context as well as the values and preferences of the patients because the evidence alone is never sufficient to inform a clinical decision.

References

1. Guyatt GH, Haynes RB, Jaeschke RZ, et al. 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(10):1290–6.
2. Guyatt G, Rennie D, Meade MO, Cook DJ. User's guide to the medical literature: a manual for evidence-based clinical practice. 2nd ed. London: AMA; 2002.
3. Delcourt R, Vastesaeger M. Action of atromid on total and beta-cholesterol. J Atheroscler Res. 1963;3:533–7.
4. A co-operative trial in the primary prevention of ischaemic heart disease using clofibrate. Report from the Committee of Principal Investigators. Br Heart J. 1978;40(10):1069–118.
5. de Bittencourt PR, Gomes-da-Silva MM. Multiple sclerosis: long-term remission after a high dose of cyclophosphamide. Acta Neurol Scand. 2005;111(3):195–8.
6. Gobbini MI, Smith ME, Richert ND, Frank JA, McFarland HF. Effect of open label pulse cyclophosphamide therapy on MRI measures of disease activity in five patients with refractory relapsing-remitting multiple sclerosis. J Neuroimmunol. 1999;99(1):142–9.
7. Marrie RA, Wolfson C, Sturkenboom MC, et al. Multiple sclerosis and antecedent infections: a case-control study. Neurology. 2000;54(12):2307–10.
8. Putzki N, Katsarava Z, Vago S, Diener HC, Limmroth V. Prevalence and severity of multiple-sclerosis-associated fatigue in treated and untreated patients. Eur Neurol. 2008;59(3–4): 136–42.
9. Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965;58:295–300.
10. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med. 2000;342(25):1878–86.
11. Taylor DW, Barnett HJ, Haynes RB, et al. Low-dose and high-dose acetylsalicylic acid for patients undergoing carotid endarterectomy: a randomised controlled trial. ASA and Carotid Endarterectomy (ACE) Trial Collaborators. Lancet. 1999;353(9171):2179–84.

12. Hankey GJ, Todd JA, Yap PL, Warlow CP. An "n of 1" trial of intravenous immunoglobulin treatment for chronic inflammatory demyelinating polyneuropathy. J Neurol Neurosurg Psychiatry. 1994;57(9):1137.
13. Popp AJ, Shafer N. Extracranial to intracranial vascular anastomosis for occlusive cerebrovascular disease: experience in 110 patients. Surgery. 1977;82(5):648–54.
14. Failure of extracranial-intracranial arterial bypass to reduce the risk of ischemic stroke. Results of an international randomized trial. The EC/IC Bypass Study Group. N Engl J Med. 1985;313(19):1191–200.
15. Montori VM, Guyatt GH. Intention-to-treat principle. CMAJ. 2001;165(10):1339–41.
16. The Canadian cooperative trial of cyclophosphamide and plasma exchange in progressive multiple sclerosis. The Canadian Cooperative Multiple Sclerosis Study Group. Lancet. 1991;337(8739): 441–6.
17. La Mantia L, Milanese C, Mascoli N, D'Amico R, Weinstock-Guttman B. Cyclophosphamide for multiple sclerosis. Cochrane Database Syst Rev. 2007(1):CD002819.
18. Brook RH, Chassin MR, Fink A, Solomon DH, Kosecoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. Int J Technol Assess Health Care. 1986;2(1):53–63.
19. Campbell SM, Braspenning J, Hutchinson A, Marshall M. Research methods used in developing and applying quality indicators in primary care. Qual Saf Health Care. 2002;11(4):358–64.
20. The periodic health examination. Canadian Task Force on the Periodic Health Examination. Can Med Assoc J. 1979;121(9):1193–254.
21. 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 Serv Res. 2004;4(1):38.
22. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20(3 Suppl):21–35.
23. Woolf SH, Battista RN, Anderson GM, Logan AG, Wang E. Assessing the clinical effectiveness of preventive maneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. A report by the Canadian Task Force on the Periodic Health Examination. J Clin Epidemiol. 1990;43(9):891–905.
24. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1989;95(2 Suppl):2S–4.
25. Guyatt G, Schunemann H, Cook D, Jaeschke R, Pauker S, Bucher H. Grades of recommendation for antithrombotic agents. Chest. 2001;119(1 Suppl):3S–7.
26. Oxford Centre for Evidence-based Medicine Levels of Evidence. http://www.cebm.net/index.aspx?o=1025. Accessed Feb 2009.
27. Edlund W, Gronseth G, So Y, Franklin G. Clinical practice guideline process manual, 2004 Edition, Appendix 9. The American Academy of Neurology. http://www.aan.com/globals/axon/assets/3749.pdf. Accessed Feb 2009.
28. Guyatt G. Evidence-based medicine. ACP J Club (Ann Intern Med). 1991;114(Suppl 2):A-16.