GUIDELINE DEVELOPMENT METHODOLOGY

Endorsed by the American Association of Neurological Surgeons (AANS), the Congress of Neurological Surgeons (CNS), and the AANS/CNS Joint Guideline Committee

The AANS, CNS, and JGC recommend that clinical practice parameter (i.e., guideline) development rely on the following methodology based on previously published AANS and CNS endorsed evidence-based reviews[1]:

Literature Search

Extensive literature searches should be undertaken for each clinical question addressed. At a minimum, the searches should include the available English-language literature for a length of time appropriate to the subject using the computerized database of the National Library of Medicine. This will vary depending on whether the guideline is an original project or an update of a previous version. Human studies are looked for, and the search terms employed should reflect the clinical question in as much detail as is relevant. Abstracts are reviewed and clearly relevant articles are selected for evaluation.

Evaluating Strength of the Therapy Literature

Each paper found by the above-mentioned techniques should be evaluated according to study type (e.g., therapy, diagnosis, clinical assessment). For therapy, evidence can be generated by any number of study designs. The strongest study protocol, when well designed and executed, is by far the randomized controlled trial (RCT). The prospectiveness, presence of contemporaneous comparison groups, and adherence to strict protocols observed in the RCT diminish sources of systematic error (i.e., bias). The randomization process reduces the influence of unknown aspects of the patient population that might affect the outcome.

The next strongest study designs are the non-randomized cohort study and the case-control study, also comparing groups who received specific treatments, but in a non-randomized fashion. In the former study design, an established protocol for patient treatment is followed and groups are compared in a prospective manner, provided their allocation to treatment is not determined by characteristics that would preclude them from receiving either treatment being studied. These groups would have a disorder of interest (e.g., spinal cord injury), receive different interventions, and differences in outcome would be studied. In the case-control study, the study is designed with the patients divided by outcome (e.g., functional ability) and with their treatment (e.g., surgery vs. no surgery) being evaluated for a relationship. These studies are more open to systematic and random error and thus are less compelling than a RCT. However, a RCT with significant design flaws that threaten its validity loses its strength and may be classified as a weaker study.

The methodological quality of RCTs and the risk of bias should be assessed using the following six criteria [2]:

1. **Sequence generation** (Was the allocation sequence adequately generated? Yes, No or Unclear)

2. **Allocation concealment** (Was allocation adequately concealed such that it could not be foretold? Yes, No or Unclear)

3. **Blinding** (Were participants, treatment providers and/or outcome assessors blinded to the treatment allocations? Yes, No or Unclear)

4. **Incomplete reporting of data** (Were incomplete outcome data adequately addressed? Yes, No or Unclear)

5. **Selective reporting of outcomes** (Were all the outcomes specified reported? Yes, No or Unclear)

6. **Other potential threats to validity** (Was the RCT free of other issues that could put it at a high risk of bias? Yes, No or Unclear)

The least strong evidence is generated by published series of patients all with the same or similar disorder followed for outcome, but not compared as to treatment. In this same category, the case report, expert opinion, and the RCT is so significantly flawed that the conclusions are uncertain. The aforementioned statements regarding study strength refer to studies of treatment. However, patient management includes not only treatment, but also diagnosis and clinical assessment. These aspects of patient care require clinical studies that are different in design and that generate evidence regarding choices of diagnostic tests and clinical measurement.

Evaluating Strength of the Diagnostic Test Literature

To be useful, diagnostic tests have to be reliable and valid. Reliability refers to the test's stability in repeated use and in the same circumstance. Validity describes the extent to which the test reflects the "true" state of affairs, as measured by some "gold standard" reference test. Accuracy reflects the test's ability to determine who does and does not have the suspected or potential disorder. Overall, the test must be accurate in picking out the true positives and true negatives, with the lowest possible false positive and false negative rate. These attributes are represented by sensitivity, specificity, positive predictive value, and negative predictive value. These may be calculated using a Bayesian 2 X 2 table as follows:

		GOLD STANDARD		
		Patient has injury	Patient has no injury	
TEST RESULT: C-SPINE FILM	Positive: Appears to have injury	TRUE POSITIVE (a)	FALSE POSITIVE (b)	(a) + (b)
	Negative: Appears to have no injury	FALSE NEGATIVE (c)	TRUE NEGATIVE (d)	(c) + (d)
		(a) + (c)	(b) + (d)	(a) + (b) + (c) + (d)

Using the above table, the components of accuracy can be expressed and calculated as follows:

Sensitivity **a/a+c**
Specificity **d/b+d**
Positive predictive value **a/a+b**
Negative predictive value **d/c+d**
Accuracy **a+d/a+b+c+d**

When considering diagnostic tests, these attributes do not always rise together. But generally speaking, these numbers should be greater than 70% to consider the test useful. The issue of reliability of the test will be discussed below when describing patient assessment.

Evaluating Strength of the Patient Assessment Literature

There are two points in the patient management paradigm when patient assessment is key. There is the initial assessment (e.g., patient's condition in the trauma room), and the ultimate, or outcome, assessment. All patient assessment tools, whether they are radiographic, laboratory or clinical, require that the measurement be reliable. In the case of studies carried out by mechanical or electronic equipment, these devices must be calibrated regularly to assure reliability. In the instance of assessments carried out by observers, reliability is assured by verifying agreement between various observers carrying out the same assessment, and also by the same observer at different times. Because a certain amount of agreement between observers or observations could be expected to occur by chance alone, a statistic has been developed to measure the agreement between observations or observers beyond chance. This is known as an index of concordance and is called the *kappa* statistic, or simply *kappa* (3). Once again, the Bayesian 2 X 2 table can be utilized to understand and to calculate kappa.

		OBSERVER #1		
		YES	NO	
OBSERVER	YES	AGREE (a)	DISAGREE (b)	(a) + (b) = f_1
#2	NO	DISAGREE (c)	AGREE (d)	(c) + (d) = f_2
		(a) + (c) = n_1	(b) + (d) = n_2	(a) + (b) + (c) + (d) = N

Using these numbers, the formula for calculating kappa is:
$k = N(a+d) - (n_1f_1 + n_2f_2) \div N^2 - (n_1f_1 + n_2f_2)$ or **$k = 2(ad - bc) \div n_1f_2 + n_2f_1$**

To translate the numbers generated by these formulas to meaningful interpretations of the strength of the agreement between observers or observations, the following guidelines are used [1]:

Value of k	**Strength of Agreement**
<0	Poor
0 - .20	Slight
.21 - .40	Fair
.41 - .60	Moderate
.61 - .80	Substantial
.81 – 1.00	Almost perfect

The above methodology applies to dichotomous variables. However, many patient assessment tools are not dichotomous and are instead ordinal or interval scales. In those cases, weighted kappas or intraclass correlation coefficients are used as the measure of reliability.

Each paper on clinical assessment should be examined for its adherence to the rules of reliability, and the kappa, weighted kappa, intraclass correlation coefficients (or similar measure) should be clearly reported and linked to the strength of recommendations, as described below.

Linking Evidence to Guidelines

The concept of linking evidence to recommendations has been further formalized by the American Medical Association (AMA) and many specialty societies, including the American Association of Neurological Surgeons (AANS), the Congress of Neurological Surgeons (CNS), and the American Academy of Neurology (AAN). This formalization involves the designation of specific relationships between the strength of evidence and the strength of recommendations to avoid ambiguity. In the paradigm for therapeutic maneuvers, evidence is classified into that which is derived from the strongest clinical studies (e.g., well-designed, randomized controlled

trials), or **Class I** evidence**. Class I** evidence is used to support recommendations of the strongest type, defined as **Level I (or A)** recommendations**,** indicating a *high degree of clinical certainty***.** Non-randomized cohort studies, randomized controlled trials with design flaws, and case-control studies (comparative studies with less strength) are designated as **Class II** evidence. These are used to support recommendations defined as **Level II (or B),** reflecting *a moderate degree of clinical certainty*. Other sources of information, including observational studies such as case series and expert opinion, as well as randomized controlled trials with flaws so serious that the conclusions of the study are truly in doubt are considered **Class III** evidence and support **Level III (or C)** recommendations, reflecting *unclear clinical certainty.* These categories of evidence are summarized in the tables below.

Classification of Evidence on Therapeutic Effectiveness

The criteria below apply to practice guidelines (parameters) for ***therapeutic effectiveness or treatment.*** One of the practical difficulties encountered in implementing this methodology is that a poorly designed randomized controlled trial might take precedence over a well-designed case-control or non-randomized cohort study. The authors of this document have attempted to avoid this pitfall by carefully evaluating the quality of the study, as well as its type.

Class I Evidence Level I (or A) Recommendation	Evidence from one or more well-designed, randomized controlled clinical trial, including overviews of such trials.
Class II Evidence Level II (or B) Recommendation	Evidence from one or more well-designed comparative clinical studies, such as non-randomized cohort studies, case-control studies, and other comparable studies, including less well-designed randomized controlled trials.
Class III Evidence Level III (or C) Recommendation	Evidence from case series, comparative studies with historical controls, case reports, and expert opinion, as well as significantly flawed randomized controlled trials.

To assess literature pertaining to ***prognosis***, ***diagnosis***, and ***clinical assessment***, completely different criteria must be used. A summary table is provided as **ATTACHMENT I,** entitled "**RATING SCHEME FOR THE STRENGTH OF THE EVIDENCE."**

Classification of Evidence on Prognosis

In order to evaluate papers addressing ***prognosis***, five technical criteria are applied:

- Was a well-defined representative sample of patients assembled at a common (usually early) point in the course of their disease?
- Was patient follow-up sufficiently long and complete?
- Were objective outcome criteria applied in a "blinded" fashion?
- If subgroups with different prognoses were identified, was there adjustment for important prognostic factors?

- If specific prognostic factors were identified, was there validation in an independent "test set" group of patients?

If all five of these criteria are satisfied, the evidence is classified as Class I. If four out of five are satisfied, the evidence is Class II, and if less than 4 are satisfied, it is Class III.

Class I Evidence Level I (or A) Recommendation	All 5 technical criteria above are satisfied.
Class II Evidence Level II (or B) Recommendation	Four of five technical criteria are satisfied.
Class III Evidence Level III (or C) Recommendation	Everything else.

Classification of Evidence on Diagnosis

For ***diagnosis***, papers are evaluated differently. The issues addressed by papers on diagnosis are related to the ability of the diagnostic test to successfully distinguish between patients who have and do not have a disease or pertinent finding. This speaks to the validity of the test and is illustrated below.

Class I Evidence Level I (or A) Recommendation	Evidence provided by one or more well-designed clinical studies of a *diverse* population using a "gold standard" reference test in a blinded evaluation appropriate for the diagnostic applications and enabling the assessment of sensitivity, specificity, positive and negative predictive values, and, where applicable, likelihood ratios.
Class II Evidence Level II (or B) Recommendation	Evidence provided by one or more well-designed clinical studies of a *restricted* population using a "gold standard" reference test in a blinded evaluation appropriate for the diagnostic applications and enabling the assessment of sensitivity, specificity, positive and negative predictive values, and, where applicable, likelihood ratios.
Class III Evidence Level III (or C) Recommendation	Evidence provided by expert opinion or studies that do not meet the criteria for the delineation of sensitivity, specificity, positive and negative predictive values, and, where applicable, likelihood ratios.

Classification of Evidence on Clinical Assessment

For clinical assessment, there needs to be both reliability and validity in the measure. This means that the assessment is done reliably between observers and by the same observer at a different time. For validity, the clinical assessment, like diagnostic tests described above, need to adequately represent the true condition of the patient. This latter aspect is difficult to measure, so most clinical assessments are graded according to their reliability.

Class I Evidence Level I (or A) Recommendation	Evidence provided by one or more well-designed clinical studies in which interobserver and/or intraobserver reliability is represented by a **Kappa statistic ≥ 0.60.**
Class II Evidence Level II (or B) Recommendation	Evidence provided by one or more well-designed clinical studies in which interobserver and/or intraobserver reliability is represented by a **Kappa statistic ≥ 0.40.**
Class III Evidence Level III (or C) Recommendation	Evidence provided by one or more well-designed clinical studies in which interobserver and/or intraobserver reliability is represented by a **Kappa statistic < 0.40.**

For each question addressed in an evidence-based report, the articles utilized in formulating the results should be referenced, summarized by study type, and assigned a classification according to the scheme outlined above. These designations should be clearly listed in **Evidentiary Tables** at the end of each document.

In every way, the author group should attempt to adhere to the Institute of Medicine (IOM) criteria for searching, assembling, evaluating, and weighting the available medical evidence and linking it to the strength of the recommendations presented in this document.

Systematic Reviews and Meta-analyses

Systematic reviews and meta-analyses are being published increasingly. A systematic review is based on a well-defined transparent search and review of all relevant publications on a given subject. The meta-analysis combines previously published data on similar studies in an attempt to assess the overall results. While no uniform methodology exists for evaluating and classifying these types of studies, in general, the Class of Evidence provided by these reports can be no better than the preponderance of the Class of Evidence in the individual papers that have been used in generating the summary.

Recommendations

The critical issue of the recommendation is that it reflect the strength of the level of evidence upon which it is based. Bias in the interpretation must be minimized in all cases. It is preferable that:

- The question being addressed is clearly stated (the more specific the better);
- The target population being addressed is clearly stated; and most importantly
- The language employed clearly indicates whether the intervention **is or is not recommended** and at what level.

Example 1:

Question
Do prophylactic anticonvulsants decrease the risk of seizure in patients with metastatic brain tumors compared with no treatment?

Target population
These recommendations apply to adults with solid brain metastases who have not experienced a seizure due to their metastatic brain disease.

Recommendation
Level 3: For adults with brain metastases who have not experienced a seizure due to their metastatic brain disease, routine prophylactic use of anticonvulsants is not recommended. Only a single underpowered randomized controlled trial (RCT), which did not detect a difference in seizure occurrence, provides evidence for decision-making purposes.

Example 2:

Question
Should patients with newly-diagnosed metastatic brain tumors undergo open surgical resection plus whole brain radiotherapy versus whole brain radiation therapy alone?

Target population
These recommendations apply to adults with newly diagnosed single brain metastases amenable to surgical resection.

Recommendations
Surgical resection plus WBRT vs. surgical resection alone
Level 1: Surgical resection followed by WBRT is recommended as a superior treatment modality in terms of improving tumor control at the original site of the metastasis and in the brain overall, when compared to surgical resection alone.

Other Methodologies for Grading the Evidence and Determining Levels of Recommendations

The JGC recognizes that other methodologies may prove to be more effective in certain situations and will consider these alternatives on an ad hoc basis. Other methodologies include:

- Agency for Healthcare Research and Quality (AHRQ): 6 levels of evidence (1A, 1B, 2A, 2B, 3, and 4) and 3 grades of recommendation (A, B, C)
- American Academy of Neurology (AAN): 4 levels of evidence (Class 1-4), 4 grades of recommendations (A,B,C & U)
- American College of Chest Physicians (ACCP) - GRADE: 3 levels of evidence, 2 grades of recommendations (endorsed by Gordon Guyatt, et al. for potential universal adoption).
- American Heart Association and American College of Cardiology (AHA/ACC): 3 levels of evidence (A-C), 4 grades of recommendations (Classes I, IIa, IIb, and III)

- North American Spine Society (NASS): 5 levels of evidence (I-V), 4 grades of recommendations (A, B, C, & I)
- Oxford Centre for Evidence-Based Medicine: 10 levels of evidence (1A-C, 2A-C, 3A, 3B, 4 & 5), 4 grades of recommendations
- US Preventive Services Task Force: 5 levels of evidence (I, II.1-3, & III), 5 grades of recommendations (A-E)

ATTACHMENT 1

RATING SCHEME FOR THE STRENGTH OF THE EVIDENCE[1]

Levels of Evidence for Primary Research Question[1]

Types of Studies

	Therapeutic Studies – Investigating the results of treatment	Prognostic Studies – Investigating the effect of a patient characteristic on the outcome of disease	Diagnostic Studies – Investigating a diagnostic test	Economic and Decision Analyses – Developing an economic or decision model
Class I	• High quality randomized trial with statistically significant difference or no statistically significant difference but narrow confidence intervals • Systematic review[2] of Class I RCTs (and study results were homogenous[3])	• High quality prospective study[4] (all patients were enrolled at the same point in their disease with ≥80% follow-up of enrolled patients) • Systematic review[2] of Class I studies	• Testing of previously developed diagnostic criteria on consecutive patients (with universally applied reference "gold" standard) • Systematic review[2] of Class I studies	• Sensible costs and alternatives; values obtained from many studies; with multiway sensitivity analyses • Systematic review[2] of Class I studies
Class II	• Lesser quality RCT (e.g., <80% follow-up, no blinding, or improper randomizatio n) • Prospective[4] comparative	• Retrospective[6] study • Untreated controls from an RCT • Lesser quality prospective study (e.g., patients enrolled at	• Development of diagnostic criteria on consecutive patients (with universally applied reference "gold" standard)	• Sensible costs and alternatives; values obtained from limited studies; with multiway sensitivity analyses

[1] *Modified and reviewed by Beverly Walters, MD*

	study[5] Systematic review[2] of Class II studies or Class I studies with inconsistent results • Case control study[7] • Retrospective [6] comparative study[5] • Systematic review[2] of Class II studies	different points in their disease or <80% follow-up) • Systematic review[2] of Class II studies • Case control study[7]	• Systematic review[2] of Class II studies • Study of nonconsecutive patients; without consistently applied "gold" standard • Systematic review[2] of Class III studies	• Systematic review[2] of Level II studies • Analyses based on limited alternatives and costs; and poor estimates • Systematic review[2] of Level III studies
Class III	• Case Series[8] • Expert opinion	• Case Series • Expert Opinion	• Case-control study • Poor reference standard • Expert Opinion	• Analyses with no sensitivity analyses • Expert Opinion

RCT= randomized controlled trial

[1] A complete assessment of quality of individual studies requires critical appraisal of all aspects of the study design.

[2] A combination of results from two or more prior studies.

[3] Studies provided consistent results.

[4] Study was started before the first patient enrolled.

[5] Patients treated one way (e.g., cemented hip arthroplasty) compared with a group of patients treated in another way (e.g., uncemented hip arthroplasty) at the same institution.

[6] The study was started after the first patient enrolled.

[7] Patients identified for the study based on their outcome, called "cases" (e.g., failed total arthroplasty) are compared to those who did not have outcome, called "controls" (e.g., successful total hip arthroplasty).

[8] Patients treated one way with no comparison group of patients treated in another way.

References:

1. *Methodology of guideline development.* Neurosurgery, 2002. **50**(3 Suppl): p. S2-6.

2. Higgins JPT, Green S (editors). Cochrane Handbook for systematic Reviews of Interventions Version 5.0.2 [updated September] 2009. The Cochrane Collaboration, 2009. Available from www.cochranehandbook.org.