2011 Edition

Clinical Practice Guideline Process Manual

Prepared by

Gary S. Gronseth, MD, FAAN
Laura Moses Woodroffe
Thomas S. D. Getchius

For the American Academy of Neurology (AAN) Guideline Development Subcommittee, AAN membership, and the public

Clinical Practice Guideline Process Manual

For more information contact:

American Academy of Neurology
1080 Montreal Avenue
St. Paul, MN 55116
(651) 695-1940
guidelines@aan.com

The authors thank the following for their contributions:

- Julie Cox, MFA, for copyediting of this edition
- Erin Hagen for her contributions to the formatting of this manual
- Wendy Edlund; Yuen T. So, MD, PhD, FAAN; and Gary Franklin, MD, MPH, for their work on the 2004 edition
- James C. Stevens, MD, FAAN; Michael Glantz, MD, FAAN; Richard M. Dubinsky, MD, MPH, FAAN; and Robert E. Miller, MD, for their work on the 1999 edition
- Members of the Guideline Development Subcommittee for their efforts in developing high-quality, evidence-based guidelines for the AAN membership

Guideline Development Subcommittee Members

John D. England, MD, FAAN, Chair
Cynthia L. Harden, MD, Vice Chair
Melissa Armstrong, MD
Eric J. Ashman, MD
Stephen Ashwal, MD, FAAN
Misha-Miroslav Backonja, MD
Richard L. Barbano, MD, PhD, FAAN
Michael G. Benatar, MBChB, DPhil, FAAN
Diane K. Donley, MD
Terry D. Fife, MD, FAAN
David Gloss, MD
John J. Halperin, MD, FAAN
Deborah Hirtz, MD, FAAN
Cheryl Jaigobin, MD
Andres M. Kanner, MD
Jason Lazarou, MD
Steven R. Messé, MD, FAAN
David Michelson, MD
Pushpa Narayanaswami, MBBS, DM, FAAN
Anne Louise Oaklander, MD, PhD, FAAN
Tamara M. Pringsheim, MD
Alexander D. Rae-Grant, MD
Michael I. Shevell, MD, FAAN
Theresa A. Zesiewicz, MD, FAAN

Suggested citation: AAN (American Academy of Neurology). 2011. *Clinical Practice Guideline Process Manual,* 2011 Ed. St. Paul, MN: The American Academy of Neurology.

Table of Contents

Preface

This manual provides instructions for developing evidence-based practice guidelines and related documents for the American Academy of Neurology (AAN). It is intended for members of the AAN's Guideline Development Subcommittee (GDS) and facilitators and authors of AAN guidelines. The manual is also available to anyone curious about the AAN guideline development process, including AAN members and the public.

Clinical practice guidelines (CPG) are statements that include recommendations intended to optimize patient care that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options.[1]

Although the goal of all practice guidelines is the same—to assist patients and practitioners in making health care decisions—different organizations use different methodologies to develop them. The AAN uses a strict evidence-based methodology that follows the Institute of Medicine's (IOM) standards for developing systematic reviews and CPGs.[1,2] All AAN guidelines are based upon a comprehensive review and analysis of the literature pertinent to the specific clinical circumstance. The evidence derived from this systematic review informs a panel of experts who transparently develop the conclusions and recommendations of the CPG using a formal consensus development process.

This manual is divided into four sections. The first is a brief introduction to evidence-based medicine (EBM). This section closes with the rationale for the AAN's adoption of the EBM methodology for the development of its practice recommendations.

The second section is an in-depth description of the EBM process as applied by the AAN. It describes the technical aspects of each step of the process—from developing questions to formulating recommendations.

The third section of the manual describes the logistics of AAN guideline development. It details the intricacies of guideline development—from proposing a guideline topic to formatting and writing an AAN guideline for publication.

The last section consists of appendices of supportive materials, including tools useful for the development of an AAN guideline.

This manual gives an in-depth description of the process that the AAN employs for developing practice guidelines. It necessarily introduces many statistical and methodological concepts important to the guideline development process. However, this manual does not comprehensively review these topics. The reader is referred to appendix 1 for a list of resources providing further information on statistical and methodological topics.

[1]Institute of Medicine of the National Academies. Clinical Practice Guidelines We Can Trust: Standards for Developing Trustworthy Clinical Practice Guidelines (CPGs). *http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust.aspx.* Released March 23, 2011. Accessed August 11, 2011.

[2]Institute of Medicine of the National Academies. Finding What Works in Health Care: Standards for Systematic Reviews. *http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx.* Released March 23, 2011. Accessed August 11, 2011.

EBM concepts are best introduced with a case such as the following example regarding ischemic stroke. A 55-year-old banker with a history of controlled hypertension is diagnosed with a small, left-hemispheric ischemic stroke. He has minimal post-stroke functional deficits. The usual stroke workup does not identify the specific cause. An echocardiogram shows no obvious embolic source but does demonstrate a patent foramen ovale (PFO). What is the best strategy to prevent another ischemic stroke in this patient?

Neurologists have varied and often strong opinions on the appropriate management of cryptogenic stroke patients with PFOs. Some would recommend closure of the PFO, as it is a potential source of paradoxical emboli. Others would consider the PFO incidental and unlikely to be causally related to the stroke.

DID YOU KNOW?
The Three Pillars

Evidence is only one source of knowledge clinicians use to make decisions. Other sources include established *Principles*—for example the neuroanatomic principles that enable neurologists to know precisely that a patient has a lesion in the lateral medulla just by examining the patient—and *Judgment*—the intuitive sense clinicians rely on to help them decide what to do when there is uncertainty. One of the goals of the EBM method of analysis is to distinguish explicitly between these sources of knowledge.

Introduction to Evidence-based Medicine

Some would choose antiplatelet medications for secondary stroke prevention whereas others would choose anticoagulation. Which treatment strategy is most likely to prevent another stroke?

Asking a question is the first step in the EBM process (see figure 1). To answer the PFO question, the EBM method would next require looking for strong evidence. So, what is evidence?

DID YOU KNOW?

It is important to remember that relative to AAN practice guidelines, the term *evidence* refers to information from studies of clinically important outcomes in patients with specific conditions undergoing specific interventions. Basic science studies including animal studies, though providing important information in other contexts, are not formally considered in the development of practice guidelines.

Evidence in an EBM context is information from any study of patients with the condition who are treated with the intervention of interest and are followed to determine their outcomes. Evidence that would inform our question can be gained from studies of patients with cryptogenic stroke and PFO who undergo PFO closure or other therapy and are followed to determine whether they have subsequent strokes. For finding such studies the EBM method requires comprehensive searches of online databases such as MEDLINE. The systematic literature search maximizes the chance that we will find all relevant studies.

When a study is found, we need to determine the strength of the evidence it provides. For this purpose EBM provides validated rules that determine the likelihood that an individual study accurately answers a question. Studies likely to be accurate provide strong evidence. Rating articles according to the strength of the evidence provided is especially necessary when different studies provide conflicting results. For example, some studies of patients with cryptogenic PFO stroke might suggest that closure lowers stroke risk whereas others might suggest that antiplatelet treatment is as effective as PFO closure. The study providing the strongest evidence should carry more weight.

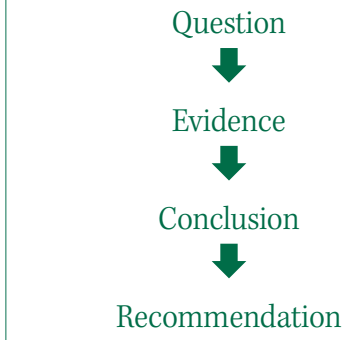
After all the relevant studies have been found and rated, the next step in the EBM process is to synthesize the evidence to answer the question. Relative to PFO, after the literature has been comprehensively searched and all the studies have been rated, one would discover that no study provides strong evidence that informs the question as to the optimal therapy. The evidence is insufficient to support or refute the effectiveness of any of the proposed treatment strategies.

When faced with insufficient evidence to answer a clinical question, clinicians have no choice but to rely on their individual judgments. The absence of strong evidence is likely one of the reasons there is such practice variation relative to the treatment of PFO. Importantly, relative to our PFO question, the EBM process tells us that these treatment decisions are judgments—that is, they are merely informed opinions. No matter how strong the opinion, no one really knows which treatment strategy is more likely to prevent another stroke.

The all-too-common clinical scenario for which there is insufficient evidence to inform our questions highlights the rationale for the AAN's decision to rely on strict EBM methods for guideline development. In the case of insufficient evidence, such as the treatment of a patient with cryptogenic stroke and PFO, an expert panel's opinion on the best course of action could be sought. This would enable the making of practice recommendations on how to treat such patients. However, endorsing expert opinion in this way would result in the AAN's substituting the judgment of its members with the judgment of the expert panel. When such opinions are discussed in an AAN guideline they are clearly labeled as opinions.

To be sure, the AAN values the opinion of experts and involves them in guideline development. However, the AAN also understands that the neurologist caring for a patient has better knowledge of that patient's values and individual circumstances. When there is uncertainty, the AAN believes decisions are best left to individual physicians and their patients after both physicians and patients have been fully informed of the limitations of the evidence.

***Figure 1.* The EBM Process**



DID YOU KNOW?
Misconceptions Regarding EBM

There are several pervasive misconceptions regarding EBM. A common one is that EBM is "cookbook medicine" that attempts to constrain physician judgment. In fact, the natural result of the application of EBM methods is to highlight the limitations of the evidence and emphasize the need for individualized physician judgment in all clinical circumstances.

EBM Process as Applied by the AAN

The EBM process used in the cryptogenic stroke and PFO scenario illustrates the flow of the EBM process (see figure 1) in the development of AAN practice guidelines. First, guideline authors identify one or more clinical question(s) that need(s) to be answered. The question(s) should address an area of quality concern, controversy, confusion, or practice variation.

Second, guideline authors identify and evaluate all pertinent evidence. A comprehensive literature search is performed. The evidence uncovered in the search is evaluated and explicitly rated on the basis of content and quality.

Third, the authors draw conclusions that synthesize and summarize the evidence to answer the clinical question(s).

Finally, the authors provide guidance to clinicians by systematically translating the conclusions of the evidence to action statements in the form of practice recommendations. The recommendations are worded and graded on the basis of the quality of supporting data and other factors, including the overall magnitude of the expected risks and benefits associated with the intervention.

The subsequent sections expand on each of these steps.

PITFALL

Many guidelines have been delayed for years because of poorly formulated questions.

DID YOU KNOW?

The first three steps of the EBM process—from question to conclusion—constitute the systematic review. If we stop at conclusions, we have not developed a guideline. Adding the additional step—from conclusions to recommendations—transforms the systematic review into a guideline.

Developing the Questions

Developing a question answerable from the evidence forms the foundation of the AAN's EBM process. The literature search strategy, evidence-rating scheme, and format of the conclusions and recommendations all flow directly from the question. Getting the questions right is critical.

Formulating an answerable clinical question is not a trivial step. It takes considerable thought and usually requires several iterations.

PICO Format

Clinical questions must have four components:

1. **Population**: The type of person (patient) involved
2. **Intervention**: The exposure of interest that the person experiences (e.g., therapy, positive test result, presence of a risk factor)
3. **Co-intervention**: An alternative type of exposure that the person could experience (e.g., no therapy, negative test result, absence of a risk factor—sometimes referred to as the control)
4. **Outcome:** The outcome(s) to be addressed

Population

The population usually consists of a group of people with a disease of interest, such as patients with Bell's palsy or patients with amyotrophic lateral sclerosis (ALS). The population of interest may also consist of patients at risk for a disease, for example patients with suspected multiple sclerosis (MS) or those at risk for stroke.

Often it is important to be very specific in defining the patient population. It may be necessary, for example, to indicate that the patient population is at a certain stage of disease (e.g., patients with *new-onset* Bell's palsy). Likewise, it may be necessary to indicate explicitly that the population of interest includes or excludes children.

DID YOU KNOW?—The PICO Format

In the EBM world the necessity of formulating well-structured clinical questions is so ingrained that there is a mnemonic in common use: PICO. This helps to remind guideline developers of the need to explicitly define all four components of a clinical question:

Some EBM gurus recommend adding two additional items to a clinical question: "T" for time, to explicitly indicate the time horizon one is interested in when observing the outcomes (e.g., disability at 3 months following a stroke); and, "S" for setting, to identify the particular setting that is the focus of the question (e.g., community outpatient setting vs. tertiary hospital inpatient setting). PICO is thus sometimes expanded to PICOTS.

Intervention

The intervention defines the treatment or diagnostic procedure being considered. The question almost always asks whether this intervention should be done. An example is, should patients with new-onset Bell's palsy be treated with steroids?

An example from the perspective of a diagnostic consideration would be: Should patients with new-onset Bell's palsy routinely receive brain imaging?

More than one intervention can be explicitly or implicitly included in the question. An example is, in patients with ALS which interventions improve sialorrhea? This more general question implies that authors will look at all potential interventions for treating sialorrhea.

It may be important to be highly specific in defining the intervention. For example, authors might indicate a specific dose of steroids for the Bell's palsy treatment of interest. Likewise, authors might choose to limit the question to steroids received within the first 3 days of palsy onset.

The way the interventions are specifically defined in the formulation of the question will determine which articles are relevant to answering the question.

Co-intervention

The co-intervention is the alternative to the intervention of interest. For therapeutic questions the co-intervention could be no treatment (or placebo) or an alternative treatment (e.g., L-3,4-dihydroxyphenylalanine [L-DOPA] vs. dopamine agonists for the initial treatment of Parkinson disease [PD]). For a population screening question, the alternative is not to screen.

The co-intervention is a bit more difficult to conceptualize for prognostic or diagnostic questions. Here the "intervention" is often something that cannot be actively controlled or altered. Rather it is the result of a diagnostic test (e.g., the presence or absence of 14-3-3 protein in the spinal fluid of a patient with suspected prion disease) or the presence or absence of a risk factor (e.g., the presence or absence of a pupillary light response at 72 hours in a patient post–cardiac arrest). Relative to a prognostic question the "co-intervention" is the alternative to the presence of a risk factor—the absence of a risk factor. Likewise, for a diagnostic test, the alternative to the "intervention"—a positive test result—is a negative test result.

Of course, there are circumstances where there may be many alternatives. The initial treatment of PD, for example, could commence with L-DOPA, a dopamine agonist or a monoamine oxidase B (MAO-B) inhibitor.

Finally, it is important to realize that there are times when the co-intervention is implied rather than explicitly stated in the question. The following is an example:

> In patients with Bell's palsy does prednisilone given with the first 3 days of onset of facial weakness improve the likelihood of complete facial functional recovery at 6 months?

Here the co-intervention is not stated but implied. The alternative to prednisilone in this question is no prednisilone.

Outcomes

The outcomes to be assessed should be clinically relevant to the patient. Indirect (or surrogate) outcome measures, such as laboratory or radiologic results, should be avoided, if doing so is feasible, because they often do not predict clinically important outcomes. Many treatments reduce the risk for a surrogate outcome but have no effect, or have harmful effects, on clinically relevant outcomes; some treatments have no effect on surrogate measures but improve clinical outcomes. In unusual circumstances—when surrogate outcomes are known to be stongly and causally linked to clinical outcomes—they can be used in developing a practice recommendation. (See the section on deductive inferences.)

When specifying outcomes it is important to specify *all* of the outcomes that are relevant to the patient population and intervention. For example, the question might deal with the efficacy of a new antiplatelet agent in preventing subsequent ischemic strokes in patients with noncardioembolic stroke. Important outcomes needing explicit consideration include the risk of subsequent ischemic stroke—both disabling and nondisabling—death, bleeding complications—both major and minor—and other potential adverse events. Every clinically relevant outcome should be specified. When there are multiple clinically important outcomes it is often helpful at the question development stage to rank the outcomes by degrees of importance. (Specifying the relative importance of outcomes will be considered again when assessing our confidence in the overall body of evidence.)

In addition to defining the outcomes that are to be measured, the clinical question should state *when* the outcomes should be measured. The interval must be clinically relevant; for chronic diseases, outcomes that are assessed after a short follow-up period may not reflect long-term outcome.

Questions should be formulated so that the four PICO elements are easily identified. The following is an example:

> **Population**: For patients with Bell's palsy
>
> **Intervention**: do oral steroids given within the first 3 days of onset
>
> **Co-intervention**: as compared with no steroids
>
> **Outcome**: improve long-term facial functional outcomes?

Types of Clinical Questions

There are several distinct subtypes of clinical questions. The differences among question types relate to whether the question is primarily of a therapeutic, prognostic, or diagnostic nature. Recognizing the different types of questions is critical to guiding the process of identifying evidence and grading its quality.

Therapeutic

The easiest type of question to conceptualize is the therapeutic question. The clinician must decide whether to use a specific treatment. The relevant outcomes of interest are the effectiveness, safety, and tolerability of the treatment. The strongest study type for determining the effectiveness of a therapeutic intervention is the masked, randomized, controlled trial (RCT).

Diagnostic and Prognostic Accuracy

There are many important questions in medicine that do not relate directly to the effectiveness of an intervention in improving outcomes. Rather than deciding to perform an intervention to treat a disease, the clinician may need to decide whether he or she should perform an intervention to determine the presence or prognosis of the disease. The relevant outcome for these questions is not the effectiveness of the intervention for improving patient outcomes. Rather, the outcome relates to improving the clinician's ability to *predict* the presence of the disease or the disease prognosis. The implication of these questions is that improving clinicians' ability to diagnose and prognosticate indirectly translates to improved patient outcomes.

For example, a question regarding prognostic accuracy could be worded, for patients with new-onset Bell's palsy, does measuring the amplitude of the facial compound motor action potential predict long-term facial outcome? The intervention of interest in this question is clearly apparent: facial nerve conduction studies. The outcome is also apparent: an improved ability to predict the patient's long-term facial functioning. Having the answer to this question would go a long way in helping clinicians to decide whether they should offer facial nerve conduction studies to their patients with Bell's palsy.

An RCT would not be the best study type for measuring the accuracy of facial nerve conduction studies for determining prognosis in Bell's palsy. Rather, the best study type would be a prospective, controlled, cohort survey of a population of patients with Bell's palsy who undergo facial nerve conduction studies early in the course of their disease and whose facial outcomes are determined in a masked fashion after a sufficiently long follow-up period.

Questions of diagnostic accuracy follow a format similar to that of prognostic accuracy questions. For example, for patients with new-onset peripheral facial palsy, does the presence of decreased taste of the anterior ipsilateral tongue accurately identify those patients with Bell's palsy? The intervention of interest is testing ipsilateral taste sensation. The outcome of interest is the presence of Bell's palsy as determined by some independent reference. (In this instance the reference standard would most likely consist of a case definition that included imaging to rule out other causes of peripheral facial palsy.)

As with questions of prognostic accuracy, the best study type to determine the accuracy of decreased taste sensation for identifying Bell's palsy would be a prospective, controlled, cohort survey of a population of patients presenting with peripheral facial weakness who all had taste sensation tested and who all were further studied to determine whether they in fact had Bell's palsy, using the independent reference standard. If such a study demonstrated that testing taste sensation was highly accurate in distinguishing patients with Bell's palsy from patients with other causes of peripheral facial weakness, we would recommend that clinicians routinely test taste in this clinical setting.

Population Screening

There is another common type of clinical question worth considering. These questions have a diagnostic flavor but are more concerned with diagnostic yield than with diagnostic accuracy. This type of question is applicable to the situation where a diagnostic intervention of established accuracy is employed. An example is, in patients with new-onset peripheral facial palsy should a physician routinely obtain a head MRI to identify sinister pathology within the temporal bone causing the facial palsy? There is no concern with regard to the diagnostic accuracy of head MRI in this situation. The diagnostic accuracy of MRI in revealing temporal bone pathology is established. The clinical question here is whether it is useful to routinely *screen* patients with facial palsy with a head MRI. The outcome of interest is the yield of the procedure: the frequency with which the MRI reveals clinically relevant abnormalities in this patient population. The implication is that if the yield were high enough, clinicians would routinely order the test.

The best evidence source to answer this question would consist of a prospective study of a population-based cohort of patients with Bell's palsy who all undergo head MRI early in the course of their disease.

Causation

Occasionally, a guideline asks a question regarding the cause-and-effect relationship of an exposure and a condition. Unlike diagnostic and prognostic accuracy questions that look merely for an association between a risk factor and an outcome, causation questions seek to determine whether an exposure causes a condition. An example is, does chronic repetitive motion cause carpal tunnel syndrome? Another example is, does natalizumab cause progressive multifocal leukoencephalopathy? The implication is that avoidance of the exposure would reduce the risk of the condition. As in these examples, causation most often relates to questions of safety.

Theoretically, as with therapeutic questions, the best evidence source for answering causation questions is the RCT. However, in many circumstances, for practical and ethical reasons an RCT cannot be done to determine causation. The outcome may be too uncommon for an RCT to be feasible. There may be no way to randomly assign patients to varying exposures. In these circumstances, the best evidence source for causation becomes a cohort survey where patients with and patients without the exposure are followed to determine whether they develop the condition. Critical to answering the question of causation in this type of study is strictly controlling for confounding differences between those exposed and those not exposed.

Determining the type of question early in guideline development is critical for directing the process. The kind of evidence needed to answer the question and the method for judging a study's risk of bias follow directly from the question type.

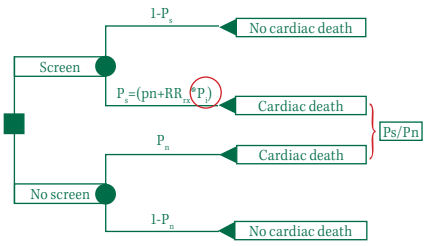
Development of an Analytic Framework

Fundamentally all CPGs attempt to answer the question, for this patient population does a specific intervention improve outcomes? The goal is to find evidence that directly links the intervention with a change in outcomes. When such direct evidence is found, it is often a straightforward exercise to develop conclusions and recommendations. When direct evidence linking the intervention to the outcome is not found, it may be necessary to explicitly develop an analytic framework to help define the types of evidence needed to link the intervention to patient relevant outcomes.

As a case in point, consider myotonic dystrophy (MD). Patients with MD are known to be at increased risk for cardiac conduction abnormalities. The question posed is, does routinely looking for cardiac problems in patients with MD decrease the risk that those patients will have heart-related complications such as sudden death? One type of analytic framework that can be constructed is a decision tree.

Figure 2 graphically depicts the factors that contribute to a decision that must be made (indicated by the black square—a decision node—at the base of the "sideways" tree). If we do not screen, the patient might or might not develop a cardiac conduction problem that leads to cardiac death (this probability is depicted by black circles—chance nodes). If we screen, the patient also has a chance of cardiac death (another chance node in figure 2), but presumably, this chance would be decreased by some degree because we have identified patients at increased risk for cardiac death and treated them appropriately (perhaps placing a pacemaker after identifying heart block on a screening EKG). The probability that screening will identify an abnormality (Pi)—conduction block on an EKG—multiplied by a measure of the effectiveness of placing a pacemaker in reducing the risk of cardiac death in patients with conduction block (RRrx) should tell us how much the risk of cardiac death is reduced with screening in patients with MD.

Figure 2. **A Decision Tree**



Direct evidence for a link between screening and reduced cardiac death would be provided by a study—ideally an RCT—that compares cardiac outcome in patients with MD who are screened with patients with MD who are not screened. If such evidence does not exist (which is probably the case) the analytic framework of the decision tree helps CPG producers identify alternative questions (and different evidence types) that might inform the decision. For example, one could find a study in which all patients with MD were

routinely screened with EKG and in which the percentage of patients with conduction block was reported. One might also find a separate study that reports the effectiveness of pacemaker placement in reducing the risk of cardiac death in patients with MD with conduction block. Using these evidence sources and the analytic framework enables a linking of the intervention and outcome.

Such analyses often suggest to guideline developers other helpful clinical questions to be asked. Rather than simply asking the therapeutic question directly linking intervention to outcome:

> For patients with MD, does routine screening with EKG (as compared with not routinely screening) reduce the risk of sudden cardiac death?

Guideline developers will also ask these questions:

> For patients with MD, how often does routine EKG screening (vs. no screening) identify patients with conduction block?
>
> For patients with MD and conduction block, does pacemaker placement (vs. no placement) reduce the risk of cardiac death?

Of course, in this example there are other potentially important outcomes to be considered, such as complications related to pacemaker screening. All important outcomes should be considered.

An analytic framework increases the likelihood that the systematic review will identify studies whose evidence, when analyzed, will answer the underlying clinical question by suggesting related questions. Additionally, the framework aids in the identification of all important outcomes.

A decision tree is one tool that is commonly used to develop an analytic framework—a causal pathway is another. Figure 3 illustrates a causal pathway used to assist in developing questions for a guideline regarding the diagnostic accuracy of tests for carpal tunnel syndrome. Regardless of the tool chosen, it is worth taking the time to use an analytic framework to help define and refine the clinical questions.

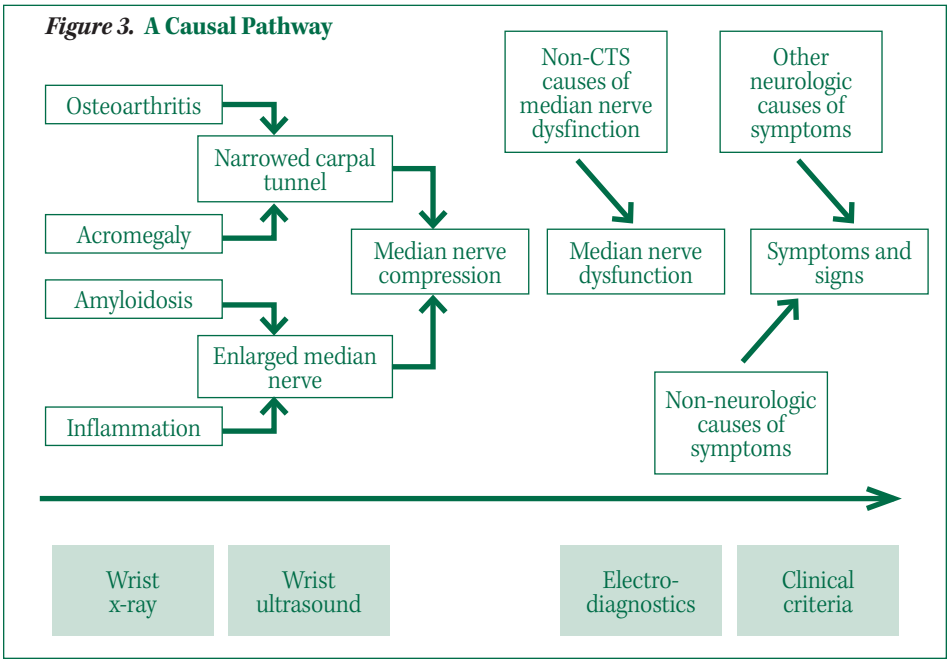
Finding and Analyzing Evidence

Finding the Relevant Evidence

A comprehensive literature search distinguishes the systematic review that forms the basis of an AAN guideline from standard review articles. The comprehensive search is performed to ensure, as much as possible, that all relevant evidence is considered. This helps to reduce the risk of bias being introduced into the process. Authors are not allowed to choose which articles they want to include (as they may select those articles that support their preconceptions). Rather, all relevant evidence is considered.

The most commonly searched database is MEDLINE. Other medical databases are also used (this is discussed further in the logistics section).

***Figure 3.* A Causal Pathway**



The initial literature search is crafted (usually with the help of a research librarian) so as to cast a wide net to ensure that relevant articles are not missed. Content experts play an important role in this step: on the basis of their knowledge of the literature they identify a few key articles they know are relevant to each of the clinical questions. These key articles are used to validate the search. If the key articles are missed in the search, the search strategy must be revised.

After completing a comprehensive search, authors use a two-step process (see figure 4) to identify relevant studies. First, authors review the titles and abstracts from the comprehensive search, to exclude citations that are obviously irrelevant to the question. Second, authors review the full text of the included titles and abstracts against prespecified inclusion and exclusion criteria. The studies meeting the inclusion and exclusion criteria form the evidence source of the guideline.

DID YOU KNOW?

Studies are included even when the guideline panel members doubt the veracity of the results. A critical assumption built into the EBM process is that investigators do not lie about or fabricate data. Unless there is direct evidence of scientific misconduct (in which case the study would likely be retracted), every study is included and analyzed using the same rules.

A secondary search of the references from review articles identified in the initial search should be made to identify any relevant studies that may have been missed.

For transparency, it is important to keep track of excluded articles and the reasons for their exclusion. After completing article selection, the authors construct a diagram depicting the flow of articles through the process, including the number excluded (see figure 5). This diagram is included in the final (published) guideline.

The identified studies meeting inclusion criteria form the evidence base that informs the review.

Identifying Methodological Characteristics of the Studies

After the studies are identified, it is necessary to extract essential characteristics of each of the studies selected for inclusion. These

extracted characteristics will be used to assess each study's strength.

The characteristics of each study will be included in a master (evidence) table. This table succinctly summarizes each study, including characteristics relevant to generalizability, risk of bias, and patient outcomes.

Elements Relevant to Generalizability

Authors should extract from the studies those elements that inform the judgment of each study's relevance to the clinical question and the generalizability of the results. These elements can be directly related to aspects of the clinical question.

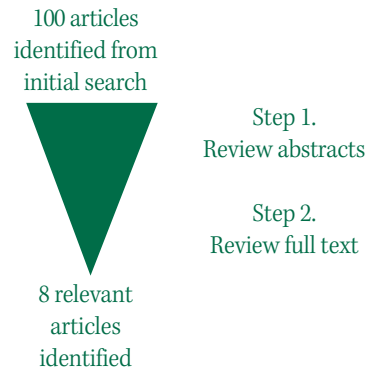
Elements relating to the patient population should include the following:

- Source of patients (e.g., neuromuscular referral center)
- Inclusion criterion used in the study to determine the presence of the condition of interest
- Age of the patients (e.g., mean and standard deviation)
- Gender of the included population (e.g., proportion female)

Elements relevant to the intervention and co-intervention should also be routinely extracted. These will be highly dependent on the clinical question but could include the following:

- Dose of medication used
- Timing of the intervention
- Nature of the diagnostic test (e.g., CT vs. MRI)

Figure 4. **Two-step Literature Review Process**



Elements relevant to the way the study measured outcomes should also be included. These will also vary from question to question but could include the following:

- Scale used to determine the outcome (e.g., global impression of change, House-Brackman vs. Adour-Swanson scale of facial function)
- Duration of follow-up

Quality-of-Evidence Indicators

Beyond the elements pertaining to generalizability, quality-of-evidence indicators should also be extracted. The items extracted will vary according to the question type.

For therapeutic questions, critical elements include the following:

- Use of a comparison (control) group
- Method of treatment allocation (randomized versus other)
- Method of allocation concealment
- Proportion of patients with complete follow-up
- Use of intent-to-treat methodologies
- Use of masking throughout the study (single-blind, double-blind, independent assessment)

For diagnostic or prognostic accuracy questions, important elements to be included are the following:

- Study design (case control versus cohort survey)
- Spectrum of patients included (narrow spectrum versus wide spectrum)
- Proportion of patients for whom both the predictor and the outcome variable are measured
- Objectiveness of the outcome variable, and whether the outcome variable is measured without knowledge of the predictor variable

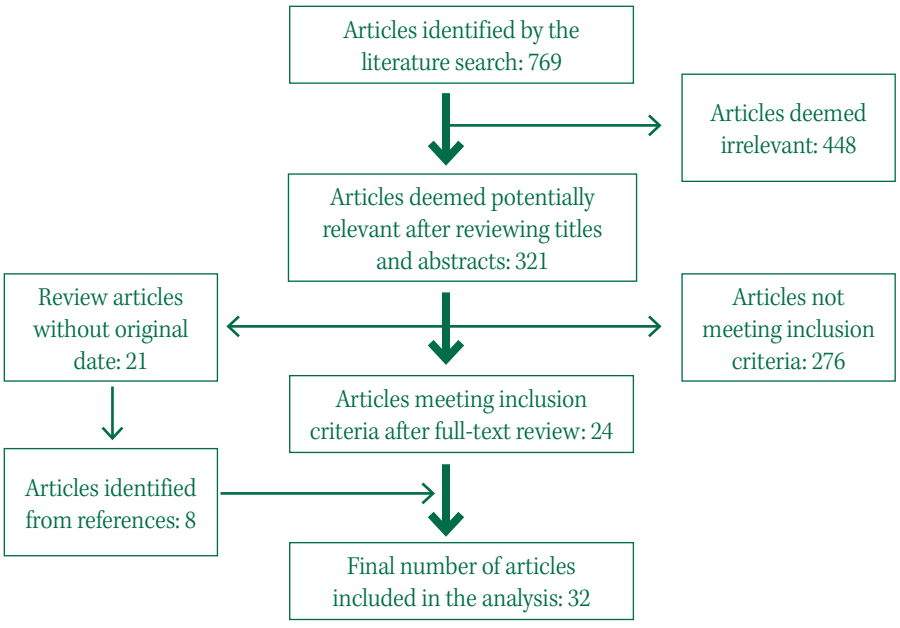
For screening questions, critical elements include the following:

- Study design (prospective vs. retrospective)
- Setting (population based, clinic based, or referral center based)
- Sampling method (selected or statistical)
- Completeness (all patients in the cohort underwent the intervention of interest)
- Masking (interpretation of the diagnostic test of interest was performed without knowledge of the patient's clinical presentation)

For causation questions, critical elements include the following:

- Study design (prospective vs. retrospective)
- Setting (population based, clinic based, or referral center based)
- Sampling method (selected or statistical)
- Completeness (all patients in the cohort underwent the intervention of interest)
- Masking (interpretation of the diagnostic test of interest was performed without knowledge of the patient's clinical presentation)
- The presence of confounding differences between those with and those without the putative causative factor

Figure 5. **Flow Diagram Documenting Disposition of Articles During the Systematic Review**



Patient Relevant Outcome Measures

Finally, patient relevant outcomes need to be extracted. These consist of a quantitative measure of what happened to patients within the study. For example, for a therapeutic question, how many patients improved? For a diagnostic question, how many patients had the disease?

Regardless of the question type, clinically relevant outcomes are usually best measured by using discrete, categorical variables rather than continuous variables. For example, the proportion of patients with Bell's palsy who have complete facial functional recovery is a more easily interpreted measure of patient outcome than the overall change in the median values of the House-Brackman facial function score.

Measuring patient outcomes using categorical variables involves counting patients. An example is, how many patients on drug X improved, and how many did not improve? Counting patients in this manner often enables construction of a contingency table. Table 1 is a simple two-by-two contingency table showing the numbers of patients improving on drug X versus placebo.

Table 1. Contingency Table

Treatment	Improved	Not Improved
Drug X	13	32
Placebo	6	78

From this it is a relatively straightforward process to calculate numeric values that express the strength of association between the intervention and the outcome. Examples are the relative risk of a poor outcome in treated patients versus untreated patients (the proportion of treated patients with a poor outcome divided by the proportion of untreated patients with a poor outcome) or the poor-outcome risk difference (the proportion of treated patients with a poor outcome minus the proportion of untreated patients with a poor outcome).

Two-by-two contingency tables can also be constructed for nontherapeutic studies. For studies regarding prognosis and causation relative risks and risk differences can also be calculated. Rather than grouping patients according to whether they received treatment, patients are grouped according to whether they had the risk factor of interest.

Quantitative measures of diagnostic accuracy can also be derived from a contingency table. These include sensitivities and specificities as well as positive and negative predictive value and likelihood ratios.

Finally, the quantitative measure used to describe the population screening studies is simply the yield, that is, the proportion of patients with the condition who are undergoing the test of interest.

Sometimes authors of the studies being considered might not report patient outcomes using categorical outcome variables. In such circumstances, if sufficient information is provided, panel members themselves should attempt to construct contingency tables. If contingency tables cannot be constructed, panel members should report the quantitative outcome measure(s) as reported in the original studies. Guideline authors are encouraged to make these determinations with the help of the facilitator or the methodological experts on the GDS.

Rating the Risk of Bias

An important step in guideline development is to measure the risk of bias in each included study. Bias, or systematic error, is the study's tendency to measure the intervention's effect on the outcome inaccurately. It is not possible to measure the bias of a study directly. (If it were, it would imply we already knew the answer to the clinical question.) However, using well-established principles of good study design, we can estimate a study's *risk* of bias.

For AAN guidelines, the risk of bias in studies is measured using a four-tiered classification scheme (see appendices 3 and 4). In this scheme, studies graded Class I are judged to have a low risk of bias, studies graded Class II are judged to have a moderate risk of bias, studies graded Class III are judged to have a moderately high risk of bias, and studies graded Class IV are judged to have a very high risk of bias. The classification rating is also known as the level of evidence.

TIP

Appendix 2 provides fomulas for calculating commonly used measures of association such as the relative risk. Additionaly, the companion spreadsheet will calculate this for you and is available at *www.aan.com/guidelines.*

Panel members assign each study a classification on the basis of that study's extracted quality-of-evidence characteristics.

The classification scheme the AAN employs accounts only for systematic error. Random error (low study power) is dealt with separately.

A study's risk of bias can be judged only relative to a specific clinical question. The standards that are applied vary among the different question types: therapeutic, diagnostic or prognostic accuracy, screening, and causation.

Appendix 4 describes in paragraph form the study characteristics needed to attain the various risk-of-bias grades. The next five sections explain in more detail each study characteristic (or element) that contributes to a study's final classification for each of the five study types (therapeutic, diagnostic, prognostic, screening, and causation).

Classifying Evidence for Therapeutic Questions

Important elements for classifying the risk of bias in therapeutic articles are described below.

Comparison (Control) Group

A comparison—or control—group in a therapeutic study consists of a group of patients who did not receive the treatment of interest. Studies without a comparison group are judged to have a high risk of bias and are graded Class IV.

To be graded Class I or Class II, studies should use concurrent controls. Studies using nonconcurrent controls, such as those using patients as their own controls (e.g., a before-after design) or those using external controls, are graded Class III.

DID YOU KNOW?

Sometimes a study provides evidence relevant to more than one question. Often in these circumstances the study will have different ratings. For example, a study could be rated Class I for a therapeutic question and Class III for a separate, prognostic question.

Treatment Allocation

To reduce the risk of bias, authors of a therapeutic article must ensure that treated and untreated patient groups are similar in every way except for the intervention of interest. In other words, known and unknown confounding differences between the treated and untreated groups must be minimized.

Randomized allocation to treatment and comparison groups is the best way to

minimize these confounding differences. Thus, to be graded Class I, a therapeutic study should have randomly allocated patients.

DID YOU KNOW?

The effect of allocation concealment on a study's accuracy has been well established. As it happens, poor allocation concealment introduces more bias into a study than failure to mask outcome assessment.

An important study characteristic that ensures patients are truly randomly allocated to different strategies is concealed allocation. Concealed allocation prevents investigators from manipulating treatment assignment. Examples of concealed allocation include use of consecutively numbered, sealed, opaque envelopes containing a predetermined, random sequence for treatment assignment and use of an independent center that an investigator contacts to obtain the treatment assignment. By comparison, examples of unconcealed allocation include flipping a coin (e.g., heads = treatment A, tails = treatment B) and assigning patients to treatment categories on the basis of the date (e.g., treatment A on odd-numbered days, treatment B on even-numbered days). These unconcealed allocation methods can be easily manipulated to control treatment allocation. For example the coin can be flipped again, or the patient can be told to come back the next day.

In addition to description of concealed allocation, Class I rating requires that panel members ensure that the randomization scheme effectively balanced the treatment and comparison groups for important confounding baseline differences. In most studies the important characteristics of each treatment group are summarized in a table (usually the first table in an article describing an RCT). If important baseline differences exist, any differences in outcomes between the different treatment groups might be explained by these baseline differences rather than by any treatment effect

Occasionally, panel members will encounter an article in which investigators attempt to match each treated patient with an untreated, comparison patient with similar baseline characteristics rather than randomly assign patients to treatment or comparison groups. Such matched studies are graded Class II.

Completeness of Follow-up

Patients enrolled in studies are sometimes lost to follow-up. Such losses occur for nonrandom reasons and may introduce confounding differences between the treated and untreated groups. Thus, Class I rating requires that more than 80% of patients within the study have completed follow-up.

For various reasons, sometimes patients initially assigned to the treatment group do not receive treatment, and patients assigned to the comparison group receive treatment. If patients cross over from the treated group to the comparison group or from the comparison group to the treated group, confounding differences can be introduced. When this happens, it is important that the investigators analyze the results using intent-to-treat principles. Put simply, such principles entail analysis of the results on the basis of whichever group (treatment or comparison) to which each patient was originally assigned.

DID YOU KNOW?

The selection of an 80% completion rate is an arbitrary one. This measure of a study's quality is best understood when positioned on a continuum—the fewer patients lost to follow-up, the better. However, to fit a study into the ordinal Class I through IV system, a cutoff had to be selected. The 80% cutoff was suggested by David Sackett, OC, FRSC—a pioneer of EBM.[1]

Masking

For a study to be graded Class I or II, an investigator who is unaware of the patient's original treatment assignment must determine the outcome. This is termed *masked or blinded outcome assessment.*

[1]Sackett, DL, Rosenberg WMC, Muir Gray JA, Haynes RB, Richardson WS. Evidence-based medicine. BMJ 1996;312:71.

PITFALL

It is important not to confuse allocation concealment with study masking (or blinding). Allocation concealment refers only to how investigators randomize patients to different treatments. *After* patients have been randomized, masking ensures that the investigators are not aware of which treatment a patient is receiving.

For a study to be graded Class III, a study investigator who is not one of the treating providers must determine the outcome. Such independent outcome assessment, although not as effective in reducing bias as masking, nonetheless has been shown to be less bias prone than having the unmasked treating physician determine the outcome. A patient's own assessment of his or her outcome (e.g., a seizure diary or completion of a quality-of-life questionnaire) fulfills the criteria for independent assessment.

The requirement for masked or independent assessment can be waived if the outcome measure is objective. An objective outcome is one that is unlikely to be affected by observer expectation bias (e.g., patient survival or a laboratory assay). Oftentimes determining whether an outcome is objective requires some judgment by the panel members. The final determination of objectiveness of any outcome is made by the AAN GDS.

Active Control Equivalence and Noninferiority Trials

Some therapeutic studies compare the efficacy of a new treatment with that of another standard treatment rather than placebo. Additional requirements are imposed on these trials.

To ensure that the new drug is being compared with an efficacious drug, there must be a previous Class I placebo-controlled trial establishing efficacy of the standard treatment.

Additionally, the standard treatment must be used in a manner that is substantially similar to that used in previous studies (Class I placebo-controlled trial) establishing efficacy of the standard treatment (e.g., for a drug, the mode of administration, dose, and dosage adjustments are similar to those previously shown to be effective).

Furthermore, the inclusion and exclusion criteria for patient selection and the outcomes of patients receiving the standard treatment are substantially equivalent to those of a previous Class I placebo-controlled study establishing efficacy of the standard treatment.

Finally, the interpretation of the study results is based on an observed-cases analysis.

Classifying Evidence for Diagnostic or Prognostic Accuracy Questions

The following paragraphs present important elements to be considered when classifying evidence for a diagnostic or prognostic accuracy question.

Comparison (Control) Group

To be useful, a study of prognostic or diagnostic accuracy should include patients with and patients without the disease or outcome of interest. Quantitative measures of accuracy cannot be calculated from studies

without a comparison group. Studies lacking a comparison group are judged to have a high risk of bias and are graded Class IV.

Study Design

A Class I study of diagnostic or prognostic accuracy would be a prospective cohort survey. Investigators would start with a group of patients suspected of having a disease (the cohort). The diagnostic test would be performed on this cohort. Some patients in the cohort would have positive test results, others negative test results. The actual presence or absence of the disease in the cohort would be determined by an independent reference standard (the gold standard). Quantitative measures of the diagnostic accuracy of the test (or predictor), such as the sensitivity or specificity, could then be calculated.

In studies of diagnostic accuracy, the steps that are followed in prognostic accuracy studies are often performed in reverse. Investigators do not start with a group of patients suspected of having the disease; rather, they select a group of patients who clearly have the disease (cases) and a group of patients who clearly do not (control). The test is then performed on both cases and controls, and measures of diagnostic accuracy are calculated. Although such case control studies are often easier to execute than cohort studies, this design introduces several potential biases. Thus, at best, such studies can be graded only Class II.

DID YOU KNOW?

Outcome objectiveness can be ranked into three tiers:

Level One: The unmasked investigator and unmasked patient cannot influence the measurement of the outcome (e.g., death, missing body part, serum glucose level).

Level Two: Either the unmasked investigator or the unmasked patient (but not both) can influence the measurement of the outcome (e.g., unmasked investigator: blood pressure measurement, MRI lesion burden; unmasked patient: seizure diary, caretaker assessment).

Level Three: Both the unmasked patient and the unmasked investigator could influence the measurement of the outcome (e.g., Unified Parkinson's Disease Rating Scale [UPDRS] score, visual analog scale score, scoring seizure scale score).

For AAN guidelines, usually only those measures meeting Level One criteria are considered objective.

PITFALL

The term *case control* is commonly misinterpreted. Many studies have "controls." The term *case control study,* however, is reserved specifically for studies wherein investigators select patients because they have the outcome of interest (e.g., the disease) or because they do not have the outcome of interest. The former are the cases; the latter are the controls.

Data Collection

For a cohort study data collection can be prospective or retrospective. In a prospective cohort study both data collection and the study itself begin before any of the patients has experienced the outcome. In a retrospective cohort study, both data collection and the study itself start after some or all of the patients have attained the outcome of interest. Retrospective data collection introduces potential bias because the investigators usually have to rely on data sources (e.g., medical records) that were not designed for the study's specific purpose. Studies with prospective data collection are eligible for a Class I rating whereas those using retrospective data collection are at best Class II.

Patient Spectrum

One of the dangers of the case control design is that such studies sometimes include either only patients who clearly have the disease or only those who clearly do not. Including such unambiguous cases can exaggerate the diagnostic accuracy of the test. To avoid this, it is important for a study employing a case control design to include a wide spectrum of patients. A wide-spectrum study would include patients with mild forms of the disease and patients with clinical conditions that could be easily confused with the disease. A narrow-spectrum study would include only patients who clearly had the disease and the control groups. Studies employing a case control design with a wide spectrum of patients can be graded Class II, and those with a narrow spectrum, Class III.

Cohort studies have a lower risk of spectrum bias than case control studies. Occasionally, spectrum bias can be introduced into a cohort study if only patients with extreme results of the diagnostic test (or risk factor) are included. For example, a study of the diagnostic accuracy of CSF 14-3-3 for prion disease would introduce spectrum bias if it included only patients with high 14-3-3 levels and patients with low 14-3-3 levels, thus excluding those with intermediate levels. The exclusion of the patients with borderline levels would tend to exaggerate the usefulness of the test.

Reference Standard

It is essential for the usability of any study of diagnostic or prognostic accuracy that a valid reference standard be used to confirm or refute the presence of the disease or outcome. This reference standard should be independent of the diagnostic test or prognostic predictor of interest. To be considered independent, the results of the diagnostic test being studied cannot be used in any way by the reference standard. The reference standard could consist of pathological, laboratory, or radiological confirmation of the presence or absence of the disease. At times, the reference standard might even consist of a consensus-based case definition. Panel members should grade as Class IV those studies that lack an independent reference standard.

Completeness

Ideally, all patients enrolled in the study should have the diagnostic test result (presence of the prognostic variable) and the true presence or absence of the disease (outcome) measured. A study is downgraded to Class II if these variables are measured for less than 80% of subjects.

Masking

For a study to be graded Class I or II, an investigator who is unaware of the results of the diagnostic test (presence or absence of the prognostic predictor) should apply the reference standard to determine the true presence of the disease (or determine the true outcome). In the instance of the case control design, for the study to obtain a Class II grade, an investigator who is unaware of the presence or absence of the disease (or unaware of the outcome) should perform the diagnostic test (measure the prognostic predictor) of interest.

For a study to be graded Class III, the diagnostic test should be performed (or prognostic predictor measured) by investigators other than the investigator who determines the true presence or absence of disease (or determines the outcome).

As with the therapeutic classification, the requirement for masked or independent assessment can be waived if the reference

standard for determining the presence of the disease (outcome) and the diagnostic test (prognostic predictor) of interest are objective. An objective measure is one that is unlikely to be affected by expectation bias.

Classifying Evidence for Population Screening Questions

For screening questions, panel members should use the study elements listed below to classify the evidence.

Data Collection

Retrospective collection of data, such as chart reviews, commonly introduces errors related to suboptimal, incomplete measurement. Thus, data collection should be prospective for a study to be classified Class I.

Setting

Studies are often performed by highly specialized centers. Because such centers tend to see more difficult and unusual cases, the patients they treat tend to be nonrepresentative of the patient population considered in the clinical question. In general, because of the potential nonrepresentativeness of patients, these studies from referral centers are graded as Class III. Occasionally, the population of interest targeted in the screening question is primarily patients referred to specialty centers. For example, some conditions that are rare or difficult to treat may be managed only at referral centers. In these circumstances, such studies can be graded Class II.

Studies of patients recruited from nonreferral centers such as primary care clinics or general neurology clinics are more representative. These studies can be graded Class II. Population-based studies tend to be the most representative and can be graded Class I.

Sampling

The ideal methods of selecting patients for a study designed to answer a screening question are selecting all patients or selecting a statistical sample of patients. Each method ensures that the patient sample is representative. Thus, a patient sample that is consecutive, random, or systematic (e.g., every other patient is included) warrants a Class I or II grade. Because patients may potentially be nonrepresentative, a study using a selective sample of patients can be graded only Class III. For example, a study of the yield of MRI in patients with headache that included patients who happened to have head MRIs ordered would be Class III because the sample is selective. A study in which MRIs are performed on all consecutive patients presenting with headache is not selective and would earn a Class I or II grade.

Completeness

For reasons similar to those given in the sampling discussion, it is important that all patients included in the cohort undergo the test of interest. If less than 80% of subjects receive the intervention of interest, the study cannot be graded higher than Class II.

Masking

For a study to be graded Class I or II for a screening question, the intervention of interest (usually a diagnostic test) should be interpreted without knowledge of the patients' clinical presentation.

Again, the requirement for independent or masked assessment can be waived if the interpretation of the diagnostic test is unlikely to be changed by expectation bias (i.e., is objective).

Classifying Evidence for Causation Questions

Particularly relative to patient safety, it may be impractical or unethical to perform RCTs to determine whether a causal relationship exists between an exposure and a disease. A classic example of this is tobacco smoking. Because of known health risks of tobacco use, no one would advocate an RCT to determine whether smoking causes lung cancer. Yet, the epidemiologic evidence for a causal relationship between smoking and cancer is overwhelming.

For such circumstances, the AAN has developed a causation evidence classification scheme. This enables investigators to assess the risk of bias of studies when the primary question is one of causation and the conduction of RCTs is not feasible.

The causation classification of evidence scheme is quite similar to the prognostic classification scheme. The former places additional emphasis on controlling for confounding differences between exposed and unexposed people. Additionally, minimal thresholds for effect size are prespecified in order for studies to qualify for Class I or II designation. Finally, nonmethodological criteria centering on biologic plausibility are included.

Making Modifications to the Classification of Evidence Schemes

The classification of evidence schemes described above provide general guidance for rating a study's risk of bias relative to a specific clinical question. These general schemes cannot identify all of the potential elements that contribute to bias in all situations. In specific circumstances, there can be nuances that require slight modifications to the schemes. For example, the outcome measures that are judged to be "objective" (i.e., unlikely to be affected by observer expectation bias) can vary on the basis of the exact clinical question. Those outcomes that will be considered objective, or any other modification to the classification of evidence schemes, need to be enumerated before study selection and data abstraction commence. This *a priori* designation of modifications is necessary to reduce the risk of bias being introduced into the review. It is acceptable to modify the classification schemes slightly to fit the specific clinical questions. However, the schemes should not be modified to fit the evidence.

Understanding Measures of Association

Interpreting the importance of the results of a study requires a quantitative measure of the strength of the association between the intervention and the outcome.

For a therapeutic question, quantitative outcomes in the treated population are usually measured relative to an untreated population. The variables used to quantitatively represent the effectiveness of an intervention are termed *measures of effectiveness* or *measures of association.* Common measures of effectiveness were introduced in the section describing study extraction and include the relative risk of an outcome (e.g., the proportion of patients with good facial outcomes in patients with Bell's palsy receiving steroids divided by the proportion of good outcomes in those not receiving steroids), or the risk difference (e.g., the proportion of patients with good facial outcomes in patients with Bell's palsy receiving steroids minus the proportion of good outcomes in those not receiving steroids.) See appendix 2 for examples of how to calculate these effect measures from contingency tables.

For articles of diagnostic or predictive accuracy, relative risks, positive and negative predictive values, likelihood ratios, and sensitivity and specificity values are the outcome variables of interest. See appendix 2 for examples of how to calculate these accuracy measures.

For screening procedures, the quantitative measure of effect will be the proportion of patients with a clinically significant abnormality identified. (See appendix 2.)

For reporting a measure of association, absolute measures are preferred (e.g., risk difference) to relative measures (e.g., relative risk). Both relative risk and risk difference are calculated from contingency tables and rely on categorical outcome measures, which are, in turn, preferred to continuous outcome measures. If the authors of the article being analyzed provide only continuous outcome measures, include these data in the evidence table.

DID YOU KNOW?

As previously mentioned, the AAN's classification of evidence scheme accounts only for the risk of bias in a study, not for the contribution of chance. Conversely, confidence intervals and *p*-values do not measure a study's risk of bias. The highest-quality study has both a low risk of bias (Class I) and sufficient precision or power to measure a clinically meaningful difference.

The mathematical tools available for measuring the contribution of chance to a study's results are much more sophisticated than our ability to measure the risk of bias.

Understanding Measures of Statistical Precision

Regardless of the clinical question type or the outcome variable chosen, it is critical that some measure of random error (i.e., the statistical power of each study) be included in the estimate of the outcome. Random error results from chance. Some patients improve and some do not regardless of the intervention used. In any given study, more patients may have improved with treatment than with placebo just because of chance. Statistical measures of precision (or power) gauge the potential contribution of chance to a study's results. In general, the larger the number of patients included in a study, the smaller the contribution of chance to the results.

Including 95% confidence intervals of the outcome measure of interest is usually the best way of gauging the contribution of chance to a study's results. A practical view of confidence intervals is that they show you where you can expect the study results to be if the study were repeated. Most of the time the results would fall somewhere between the upper and lower limits of the confidence interval. In other words, on the basis of chance alone, the study results can be considered to be consistent with any result within the confidence interval.

The *p*-value is the next best measure of the potential for random error in a study. The *p*-value indicates the probability that the difference in outcomes observed between groups could be explained by chance alone. Thus a *p*-value of 0.04 indicates that there is a 4% probability that the differences in outcomes between patient groups in a study are related to chance alone. By convention a *p*-value of < 0.05 (less than 5%) is usually required for a difference to be considered statistically significant.

The presence of a statistically significant association can also be determined by inspection of the upper and lower limits of the 95% confidence intervals. If the measure of association is the relative risk or odds ratio of an outcome, for example, and the confidence interval includes 1, the study does not show a statistically significant difference. This is equivalent to stating that the *p*-value is greater than 0.05.

Relative to measures of statistical precision, 95% confidence intervals are preferred over *p*-values. If *p*-values are not provided, include measures of statistical dispersion (e.g., standard deviation, standard error, interquartile range).

Interpreting a Study

Armed with the measure of association and its 95% confidence interval, we are in a position to interpret a study's results. Often the temptation here is to determine merely whether the study was positive (i.e., showed a statistically significant association between the intervention and outcome) or negative (did not show a statistically significant association). In interpreting study results, however, four, not two, outcomes are possible. This derives from the fact that there are two kinds of differences you are looking for: whether the difference is statistically significant and whether the difference is clinically important. Henceforth when we use the term *significant* we mean *statistically significant*, and when we use the term *important* we mean *clinically important.* From these two types of differences, four possible outcomes can be seen:

1. The study showed a significant and important difference between groups.

   For example an RCT of patients with cryptogenic stroke with PFO demonstrates that 10% of patients who had their PFO closed had strokes whereas 20% of patients who did not have their PFO closed had strokes (risk difference 10%, 95% confidence intervals 5%–15%). This difference is statistically significant (the confidence intervals of the risk difference do not include 0) and clinically important (no one would argue that a finding of 10% fewer strokes is unimportant).

2. The study showed a significant but unimportant difference between groups.

   A separate RCT enrolling a large number of patients with cryptogenic stroke with PFO demonstrates that 10.0% of patients who had their PFO closed had strokes whereas 10.1% of patients who did not have their PFO closed had strokes (risk difference 0.1%, 95% confidence intervals 0.05%–0.015%). This difference is statistically significant but arguably not clinically important (there are only 1 in 1000 fewer strokes in the patients with PFO closure).

3. The study showed no significant difference between groups, and the confidence interval was sufficiently narrow to exclude an important difference.

   A third RCT enrolling a large number of patients with cryptogenic stroke with PFO demonstrates that 5% of patients who had their PFO closed had strokes whereas 5% of patients who did not have their PFO closed had strokes (risk difference 0%, 95% confidence intervals -0.015%–0.015%). This difference is not statistically significant. Additionally the 95% confidence intervals are sufficiently narrow to allow us to confidently exclude a clinically important effect of PFO closure.

DID YOU KNOW?

The *Neurology*® journal editorial policy prohibits use of the term *statistically significant* in manuscript submissions. Instead authors are advised to use the term *significant* to convey this statistical concept. For more information on the journal's editorial policy, visit *http://submit.neurology.org.*

4. The study showed no significant difference between groups, but the confidence interval was too wide to exclude an important difference.

   Our last hypothetical RCT of patients with cryptogenic stroke with PFO demonstrates that 5% of patients who had their PFO closed had strokes whereas 5% of patients who did not have their PFO closed had strokes (risk difference 0%, 95% confidence intervals -10%–10%). This difference is not statistically significant. However, the 95% confidence intervals are too wide to allow us to confidently exclude a clinically important effect of PFO closure. Because of the lack of statistical precision, the study is potentially consistent with an absolute increase or decrease in the risk of stroke of 10%. Most would agree that a 10% stroke reduction is clinically meaningful and important.

Let us consider these outcomes one at a time.

Scenario 1 represents the clearly positive study and scenario 3 the clearly negative study. A Class I study pertinent to scenario 1 or 3 would best be described as an *adequately powered* Class I study.

Scenario 2 usually results from a large study. The study has a very high degree of power and can show even minor differences. The minor differences may not be important. The study should be interpreted as showing no meaningful difference. A Class I study pertinent to scenario 2 would best be described as an *adequately powered* Class I study showing no important difference.

Scenario 4 results from a small study. The study is so underpowered that it is unable to show significant differences even when there might be important differences. It would be inappropriate to interpret this study as negative. A Class I study pertinent to scenario 4 should be described as an *inadequately powered* Class I study.

To be sure, determining what is clinically important involves some judgment. Discussion among panel members will often resolve any uncertainty. When the clinical importance of an effect remains uncertain, it is best to stipulate explicitly in the guideline what you considered clinically important.

The methodological characteristics of each informative study along with their results should be summarized in evidence tables. See appendix 5 for a sample evidence table.

PITFALL

A common error when interpreting a study that shows no significant difference between treatment groups is to fail to determine whether the study had adequate power to exclude a clinically important difference. Such a study is not truly negative—rather, it is inconclusive. It lacks the precision to exclude an important difference.

Synthesizing Evidence—Formulating Evidence-based Conclusions

At this step multiple papers pertinent to a question have been analyzed and summarized in an evidence table. These collective data must be synthesized into a conclusion. The goal at this point is to develop a succinct statement that summarizes the evidence in answer to the specific clinical question. Ideally, this summary statement should indicate the magnitude of the effect and the class of evidence on which it is based. The conclusion should be formatted in a way that clearly links it to the clinical question.

Four kinds of information need to be considered when formulating the conclusion:

- The class of evidence
- The measure of association
- The measure of statistical precision (i.e., the random error [the power of the study as manifested by the width of the confidence intervals])
- The consistency between studies

For example, in answer to the clinical question:

> For patients with new-onset Bell's palsy,
>
> Do oral steroids given within the first 3 days of onset
>
> Improve long-term facial outcomes?

The conclusion may read:

> For patients with new-onset Bell's palsy,
>
> Oral steroids given within the first 3 days of onset of palsy
>
> Are likely safe and effective to increase the chance of complete facial functional recovery (rate difference 12%) (two inadequately powered Class I studies and two Class II studies).

In this example, the level of evidence on which the conclusion is based is indicated in two ways: 1) the term *likely safe and effective* indicates that the effectiveness of steroids is based on moderately strong evidence, and 2) the number and class of evidence on which the conclusion is based are clearly indicated in parentheses. To avoid confusion, you should explicitly indicate in the conclusion when studies have insufficient power to exclude a meaningful difference. Appendix 6 provides guidance on translating evidence into conclusions.

The level of certainty directly relates to the *highest* class of evidence with adequate power used to develop the conclusion. Thus, conclusion language will vary on the basis of the following levels of evidence:

- Multiple Class I studies:
  Are *highly likely to be* effective…
- Multiple Class II studies or a single Class I study:
  Are *likely* effective…
- Multiple Class III studies or a single Class II study
  Are *possibly* effective…
- Multiple Class IV studies or a single Class III study:
  For patients with new-onset Bell's palsy, there is *insufficient evidence to support or refute* the effectiveness of steroids in improving facial functional outcomes.

Analogous verbiage is used when studies demonstrate that therapy is ineffective:

- Multiple negative, adequately powered Class I studies:
  Are *highly likely not to be* effective…
  Are *highly likely to be in*effective…
- Multiple negative, adequately powered Class II studies; or a single adequately powered Class I study:
  Are *likely* not effective…
  Are *likely in*effective…
- Multiple negative, adequately powered Class III studies; or a single adequately powered Class II study:
  Are *possibly* not effective…
  Are *possibly in*effective…

DID YOU KNOW?

When formulating evidence-based conclusions the AAN avoids the terms *proven* effective or *established as* effective. Evidence is never definitive, and therefore conclusions derived from evidence cannot be "proven" or definitively "established."

- Multiple Class IV studies, a single adequately powered Class III study; or negative, inadequately powered Class I, II, or III studies:

  For patients with new-onset Bell's palsy, there is *insufficient evidence to support or refute* the effectiveness of steroids in improving facial functional outcomes.

Please see appendix 6 for a tool to help you construct conclusions.

Accounting for Conflicting Evidence

When all of the studies demonstrate the same result, are of the same class, and are consistent with one another, developing the conclusion is a straightforward matter.

Often, however, this is not the case. The following provides guidance on how to address inconsistent study results.

Consider a hypothetical example where the search strategy identified one Class I study, one Class II study, and one Class III study on the effectiveness of steroids in Bell's palsy. The Class I study shows a significant and important difference from placebo. The Class II and III studies show no significant or important difference from placebo. What should the author panel do? One approach would be to treat each study like a vote. Because the majority of studies (2/3) show no benefit, the panel could conclude that steroids have no effect. This vote-counting approach is not acceptable; it ignores the sources of error within each study.

The appropriate approach to take when faced with inconsistent results in the included studies is to attempt to explain the inconsistencies. The inconsistencies can often be explained by systematic or random error.

Considering Bias First: Basing the Conclusion on the Studies with the Lowest Risk of Bias

The authors should consider systematic error first. In this example, the differences in risk of bias among the studies likely explain the inconsistencies in the results. The Class I study has a lower risk of bias than the Class II or Class III studies. Thus, the results of the Class I study are more likely to be closer to the truth. The Class II and III studies should be discounted, and, if possible, the conclusion formulated should be based solely on the Class I study.

The conclusion would be worded:

Oral steroids are *likely* effective to…

(The "likely effective" conclusion is supported when there is a *single* Class I study used to formulate the recommendation. If we changed this example slightly and included *two or more* positive Class I studies, the conclusion would read "highly likely to be effective.")

Considering Random Error: Are Some Studies Underpowered?

Consider another hypothetical example: that the search strategy identified three Class I studies on the effectiveness of steroids for Bell's palsy. Assume one study showed a significant and important benefit from steroids and two studies did not.

Systematic error does not obviously explain the difference, as all three studies are Class I. Therefore, the authors must consider the random error (statistical precision or power) of the studies by looking at the confidence intervals. If the confidence intervals of all of the studies overlap, it is likely that random error (i.e., the lack of statistical power in some of the studies) explains the difference in the studies' results. On the basis of a single adequately powered Class I study a "likely effective" conclusion would be justified.

Knowing When to Perform a Meta-analysis

Another solution in this circumstance would be to perform a meta-analysis. This increases the statistical precision of the conclusion by combining all of the studies. Meta-analysis is a technique that reduces random error (but not systematic error). In this circumstance, the combined estimate of the effect of steroids would be used to develop the conclusions. For the purpose of developing conclusions for an AAN guideline, when studies are combined in a meta-analysis to increase statistical precision, the resulting pooled data are treated as though they derived from a single study.

Combining studies in a meta-analysis is often a useful way to reduce random error. However, such a practice can be inappropriate when there are differences in study design, patient populations, or outcome measures.

The strength of the conclusion ("highly likely," "likely," or "possibly effective") would depend on the lowest level of evidence used in the meta-analysis. In this situation, Class I evidence from three studies would support using the terminology "likely effective."

Another situation in which a meta-analysis may be applicable is if all three of the Class I studies in the example were negative. In the case of consistent negative studies, it is still important to look at the potential contribution of random error before formulating a conclusion. In this case, it might be a mistake to conclude that steroids are "highly likely not effective." If the confidence intervals from the studies were wide—meaning that the confidence intervals included a potentially clinically important benefit of steroids because of a lack of statistical precision in the studies—the individual studies would be inconclusive. Combining the negative studies in a meta-analysis might increase the statistical power sufficiently (i.e., narrow the confidence intervals) so that a clinically important benefit of steroids is excluded. An appropriate negative conclusion could then be made.

Methodological experts on the committee can help authors perform a meta-analysis, if necessary.

Considering Both Bias and Random Error

Consider a final example. Here the search strategy identifies five articles looking at the effectiveness of steroids in Bell's palsy. Two studies are Class I, two studies Class II, and one study Class III. The studies are inconsistent in that the Class III study and Class II studies demonstrate a statistically significant difference, and the Class I studies do not.

The authors should first examine the studies with the lowest risk of bias—the Class I studies—for systematic error. They should next examine these same studies for random error. Although both Class I studies show no benefit of steroids, both studies are underpowered. They have wide confidence intervals that include potentially clinically important benefits of steroids. Combining them in a meta-analysis still shows no significant effect of steroids. However, the combined confidence interval is too wide to exclude a benefit.

Next the authors should examine the Class II studies by performing a meta-analysis that includes both the Class I and Class II studies. The meta-analysis shows a statistically significant benefit of steroids, so the authors can now formulate a conclusion.

The example conclusion used at the beginning of this section would be appropriate for this evidence. Because Class II evidence was used in the conclusion formulation, "likely effective" is used to indicate the level of certainty.

Understanding Reasons for Inconsistencies Aside from Systematic Error and Random Error

Inconsistencies between studies cannot always be explained by a systematic consideration of the level of evidence and random error. Sometimes differences between the study populations, interventions, and outcome measures are sufficient to explain inconsistencies. At times, the inconsistencies cannot be explained. In such instances it is best acknowledged that there is insufficient evidence to draw conclusions.

Methodological experts of the subcommittees can guide panel members in this situation.

Wording Conclusions for Nontherapeutic Questions

The examples of conclusion formulation given thus far have related to therapeutic questions. Analogous procedures are followed for questions of diagnostic or prognostic accuracy and for screening questions. The conclusions are worded slightly differently in that the term *useful* is substituted for *effective*. Thus, a conclusion regarding the prognostic accuracy of facial compound motor action potential in identifying patients at increased risk of poor facial function might read:

> For patients with new-onset Bell's palsy,
>
> The measurement of facial compound motor action potentials is likely useful
>
> To identify patients at increased risk for poor facial functional recovery (sensitivity 85%, specificity 75%) (three Class II studies).

Capturing Issues of Generalizability in the Conclusion

At times the best evidence relevant to the question posed may be limited by issues of generalizability. In such circumstances, the evidence does not exactly answer the question that was posed. Rather, it answers a relevant, closely related question. Limited generalizability can arise in situations that directly relate to the PICO elements of the posed question.

The population may not be directly representative of the entire population of interest. This can arise when the highest-class studies pertinent to a question only include a subpopulation of patients with the disease. For example, the best studies of Bell's palsy might have been performed on women and not men.

Limited generalizability can also result when all relevant studies determined only the efficacy of a narrow range of possible interventions encompassed by the question. For example, if all studies of patients with Bell's palsy were limited to prednisilone at 80 mg a day for 3 days taken within 24 hours of palsy onset (no other steroid being studied), the generalizability of this evidence to other steroids at different doses and durations is limited.

Generalizability issues can also arise relative to the comparative intervention used. For example, if the literature search found only studies showing improved outcomes in patients with Bell's palsy receiving steroids as compared with patients receiving thiamine (and not placebo), the applicability of this evidence to the question of steroids as compared with placebo is limited.

Finally, generalizability issues may arise relative to the measurement of the outcome. For example, a study of steroids in patients with Bell's palsy may have determined outcome only at 2 months. It would be difficult to generalize this evidence to long-term outcomes.

When the generalizability of the evidence is limited, the conclusion should be worded to indicate the limited applicability of the evidence. Thus if only high-quality (Class I) studies of patients with Bell's palsy examining facial functional outcomes at 2 months in women treated with prednisilone at 80 mg a day for 3 days taken within 24 hours of palsy onset as compared with those of women treated with thiamine, the conclusion should *not* read as follows:

> For patients with Bell's palsy it is highly likely that steroids (as compared with placebo) improve facial functional outcomes (risk difference 12%, 95% CI 7%–15%, multiple Class I studies).

Rather, the conclusion should be worded to capture the limited generalizability of the evidence:

> For women with Bell's palsy it is highly likely that prednisilone 80-mg daily for 3 days taken within 24 hours of palsy onset as compared with thiamine improves facial functional recovery at 2 months (risk difference 12%, 95% CI 7%–15%, multiple Class I studies).

Making Practice Recommendations

The strictly evidence-based conclusions formulated using the rules discussed in the "synthesizing evidence" section defines the end of the systematic review process. The next step in the process is to develop practice recommendations.

DID YOU KNOW?

Occasionally, after completing the systematic review, guideline developers will realize that the evidence base is too weak to support any meaningful practice recommendations. In these circumstances it is appropriate to terminate the development process rather than attempt to develop practice recommendations. The systematic review itself has value in informing neurologists and patients of the limitations of the evidence. The systematic review should be published as an AAN evidence report.

The first goal of the process of making recommendations is to develop an actionable recommendation that addresses the clinical question. For example, one question regarding patients with Bell's palsy is whether we should treat them with steroids to increase the likelihood of facial functional recovery. This includes identifying the patient population, intervention, and outcome of interest. (Here the co-intervention—no treatment—is implied.) A recommendation resulting from a review of the effectiveness of treatments for Bell's palsy might read as follows:

> For patients with new-onset Bell's palsy
>
> Clinicians should offer oral steroids within the first 3 days of palsy onset
>
> To improve facial functional outcomes.

The second goal is to determine and transparently indicate our confidence that adherence to the recommendation will improve outcomes. Confidence in the strength of a recommendation in an AAN guideline is indicated by a designation of recommendation strength of Level "A," "B," or "C." Determining the recommendation level involves much more than a consideration of the quality of evidence on which the recommendation is based.

DID YOU KNOW?

In the AAN guideline development process the term *class* is used to designate the risk of bias whereas the term *level* is used to designate the strength of a recommendation.

Attaining these goals of the recommendation development process requires three steps: first, rate our confidence in the ability of the evidence to support practice recommendations; second, place the evidence into a clinical context by explicitly considering all factors that could influence the recommendation; and finally, craft the recommendation.

Rating the Overall Confidence in the Evidence from the Perspective of Supporting Practice Recommendations

The implicit assessment of the quality of evidence signaled by the terms *possibly* or *likely* measures our confidence that an estimate of the effect of an intervention is correct.[1] If our purpose was to develop only a systematic review, we would stop here. However, when possible, we want to go further by developing actionable recommendations that provide guidance to physicians and patients. To do so requires us to take a second, higher-level look at the evidence. In this second look we are not trying to estimate our confidence in the accuracy of the evidence as it relates to the effect of an intervention. Rather, we are determining whether our confidence in the evidence is sufficient to support practice recommendations. The difference is subtle but important. This second determination requires a more nuanced consideration of the evidence. To do this the AAN has adopted a modified version of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) process.[1]

DID YOU KNOW?

In adopting the GRADE process[1] for evidence synthesis, the AAN has made one major modification. The confidence level is anchored to the designation of risk of bias of the informative studies. Other factors can upgrade our confidence in the evidence only by one level above the anchored level. This is done to avoid situations wherein highly biased evidence is used to support a "high" level of confidence in the evidence. It is thought that this modification will improve the reliability of the GRADE process.[1]

The modified GRADE process for evidence synthesis used by the AAN has several steps.

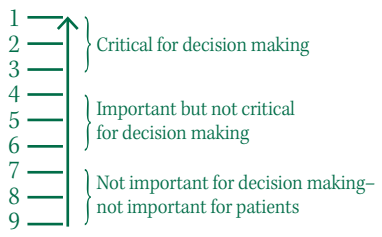
- First, we rate the clinical importance of all outcomes, designating which outcomes are critical or important.
- Subsequently, for every outcome we
  - Anchor our confidence in the evidence to the risk of bias of the best informative studies
  - Consider factors that may downgrade our confidence in the evidence
  - Consider factors that may upgrade our confidence in the evidence
  - Estimate our confidence in the evidence relative to the outcome of interest on the basis of the preceding factors
- Finally, our overall confidence in the evidence relative to all outcomes is estimated using the lowest level of confidence estimate for any critical outcome.

[1]Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol. 2011;64:380–382.

Rating the Importance of All Outcomes to Patients

Figure 6 illustrates the ordinal scale used to rank the importance of all pertinent outcomes relative to their importance to patients. Lower numbers designate higher importance outcomes in this scale. Outcomes rated 1 through 3 are considered "critical," those rated 4 through 6 "important," and those greater than 6 "not important." These ratings are made on the basis of the judgment of participants in the guideline development process. Occasionally there are published values—known as *utilities*—for which numbers are assigned to health states from 0 (deceased) to one (healthy). These utility values can assist CPG developers in designating the importance of outcomes.

Figure 6. **Hierarchy of Outcomes According to Importance to Patients**



Anchoring the Confidence Level to the Risk of Bias

A level of confidence in the evidence relative to each outcome is then assigned. Four levels of confidence are available. Although it may differ from the final designation of confidence, the confidence in the evidence is initially anchored to the class of evidence, using a method identical to that used in the development of evidence-based conclusions:

- High confidence (anchor: two Class I studies—corresponds to a "highly likely" conclusion)
- Moderate confidence (anchor: one Class I study or two Class II studies—corresponds to a "likely" conclusion)
- Low confidence (anchor: one Class II study or two Class III studies—corresponds to a "possibly" conclusion)
- Very low confidence (anchor: < two Class III studies—corresponds to an "insufficient" conclusion)

Considering Factors That Potentially Downgrade the Confidence in the Evidence

Next, explicitly consider factors that could downgrade the confidence in the evidence to support recommendations. These factors include the number and consistency of the informative studies, the statistical precision (or power) of the studies, the directness (or generalizability) of the evidence, and the presence of potential reporting bias. In general, the level of confidence in the evidence should be downgraded only by one level for these factors even when multiple factors are present.

Considering Factors That Potentially Upgrade the Confidence in the Evidence

Also explicitly consider factors that could upgrade our confidence in the evidence. Such factors include a large magnitude of effect (an effect can be so large that it likely overcomes any bias in a study), the presence of a dose response relationship (this lends plausibility to the biologic effect of the intervention), the direction of bias (if the direction of bias moves in one direction but the measured effect moves in the other, we can be more confident that the observed effect is real; this is especially true of studies that show no effect, as the direction of bias for most studies is toward showing an effect), and the biologic plausibility of the intervention (for some interventions it simply makes sense that they work; for example, in the case of

myasthenia gravis, the pathophysiology of the disease suggests that removing autoantibodies through plasma exchange should improve the condition). Importantly, these factors can result in upgrading of the confidence in the evidence by only one level.

After determining the level of confidence in the evidence for each outcome, determine the overall level of confidence in the evidence for all outcomes on the basis of the lowest confidence of the critical outcomes. For example if the evidence level for one critical outcome (e.g., death) is "high" and the corresponding confidence level for another critical outcome is "moderate," the overall confidence for all outcomes will be designated "moderate."

The level of confidence in the evidence to support recommendations derived using the modified GRADE process is one important factor to be considered when developing practice recommendations.

Putting the Evidence into a Clinical Context

Much more than evidence must be considered when crafting practice recommendations. The evidence-based conclusion and our confidence in the ability of the evidence to support practice recommendations form the foundation. Other factors influence the structure of the recommendation built on this foundation. The exact wording of the final recommendation must be carefully crafted to account for these factors. The impact of such factors varies from guideline to guideline. Below we provide some general guidance.

Developers of an AAN guideline must place the evidence into a realistic clinical context to ensure recommendations are as helpful as possible to clinicians and patients. Non–evidence-based factors that need to be transparently and systematically considered when formulating recommendations include the following:

- Deductive inferences from accepted principles
- The relative value of the benefit as compared with the risk; this is derived from a consideration of both:
  - The clinical importance of the size of the intervention's effect
  - The risk of harm of the intervention (i.e., tolerability and safety)
- The availability of the intervention
- The cost of the intervention
- The availability of potentially effective alternatives to the intervention
- The expected variation in patient preferences relative to the risks and benefits of the intervention
- The effectiveness of the intervention among important subgroups of patients with the condition

The ways in which these non–evidence-based factors may influence a recommendation will now be explored.

PITFALL

At times (some would say usually) estimates of harm from an intervention cannot be made from high-quality evidence. Low-quality sources of evidence such as case reports or registries may be the best evidence available. Such sources of evidence of harm should not be disregarded.

Understanding the Role of Deductive Inferences from Accepted Principles

There are times when linking the evidence to a recommendation requires a deductive inference from first principles. In this circumstance it is the combination of the evidence and the inference that informs the practice recommendation. A formal analytic framework such as a decision tree or causal pathway can help identify these inferences (see "Developing the Questions" section). Such inferences most commonly need to be made when a nontherapeutic question is asked. For example, let us suppose that high-quality evidence rated by the screening classification of evidence scheme indicates that a large proportion of patients with new-onset trigeminal neuralgia have potentially treatable abnormalities identified by head MRI (such as a mass of lesions compressing the trigeminal nerve). This evidence alone does not indicate that patients with trigeminal neuralgia will have better outcomes if they routinely undergo MRI. Rather, the evidence simply tells us is that a large number of such patients will have treatable abnormalities. However, if we explicitly accept a principle of care—that identifying treatable conditions is important—then in this example the inference logically follows that clinicians should routinely order head MRIs on patients with new-onset trigeminal neuralgia to identify potentially treatable causes. The axiomatic principle allows us to connect the dots from evidence to recommendation.

To ensure that the process of formulating recommendations is transparent, it is important to make such "dot-connecting" explicit. Consider a second example. Our systematic review of the evidence regarding assays of CSF 14-3-3 protein for the diagnosis of Creutzfeldt-Jakob disease (CJD) indicates that the test is 85% sensitive and 90% specific. Does the 14-3-3 test's moderately high accuracy indicate that the test is necessary for all patients with rapidly progressive dementia? No, there is more to consider. The principle of care discussed above—identifying treatable conditions is important—does not apply, as CJD is not treatable. Another principle of care could be defensibly accepted in this situation—reducing uncertainty even for untreatable conditions is important because it helps patients and families cope better with a devastating illness. It is likely that few clinicians would disagree with this principle—but some might. Explicitly stating the principles used in the formation of the recommendations serves to make the process transparent. If a person disagrees with the recommendation, the reason for the disagreement will be apparent—the person does not accept this principle as axiomatic.

In the 14-3-3 example the evidence and the explicit adoption of the principle of care do not in themselves support a recommendation to perform 14-3-3 assays routinely in patients with suspected CJD. Although of moderately high diagnostic accuracy, the 14-3-3 assay is an imperfect test. The test will not importantly change the probability of CJD in patients who are unlikely to have CJD to begin with. For example a 72-year-old with dementia progressing over 18 months is very unlikely to have CJD, and a positive 14-3-3 test is most likely to represent a false positive. Likewise, the 14-3-3 assay will provide minimal information for patients with a high likelihood of having the disease. For example a 54-year-old with rapidly progressive dementia over 3 months with symmetric diffusion-weighted-imaging changes in the basal ganglia is very likely to have CJD. A negative 14-3-3 test in this situation would most likely represent a false negative. These inferences are not derived from evidence as defined in the EBM context. Rather, they are inferred from known principles of the course of CJD and Baye's theorem (an important principle regarding contingent probabilities).

DID YOU KNOW?

In order to determine the effect of a diagnostic test on patient outcomes one must perform a utility study. Such studies involve comparing patient relevant outcomes in patients who get the test with outcomes of those who do not get the test. The utility of mammography has been tested in this way. Women were randomized either to receive routine mammography or not to receive it. In these studies outcomes (death secondary to breast cancer) were a little better in the women getting mammography. Utility studies would be rated by the AAN's therapeutic classification of evidence system rather than by the diagnostic accuracy system and could support an actionable recommendation such as one that states "should offer."

After examining the evidence and making several inferences from multiple explicitly stated principles (and assuming there are no other factors to consider), we might formulate a recommendation for the 14-3-3 example that reads something like this:

For patients with rapidly progressive dementia who are strongly suspected of having CJD and for whom diagnostic uncertainty remains, clinicians should order CSF 14-3-3 assays to reduce the uncertainty of the diagnosis.

When crafting recommendations guideline developers must consider explicitly and enumerate any principle-based inferences that support the recommendations. Additionally, the strength of the inference must be considered. Not all principle-based inferences are convincing. The author panel and oversight guideline committee will determine how compelling the inferences are, using a modified Delphi process.

Unanimously accepted principles will be labeled *compelling* and can support a Level A recommendation (assuming that the confidence in the evidence used in the inference is high). Those inferences accepted by more than 80% of the author panel and GDS will be labeled *convincing*. Convincing inferences can support at most a Level B recommendation (assuming our confidence in the evidence is at least low). Inferences accepted by more than 50% but less than 80% of participants are labeled *plausible*. Plausible inferences can support at best a Level C recommendation (assuming there is at least weak evidence). Inferences accepted by less than 50% of participants are labeled *unconvincing* and cannot support any recommendation. Regardless of the strength of the accepted principles, inferences based on insufficient evidence (i.e., very low confidence) do not usually support recommendations.

As previously illustrated, inferences from principles are most often used in conjunction with evidence to develop recommendations. There are unusual circumstances where compelling inferences alone can support practice recommendations without evidence. Recommendations based on compelling inferences from first principles are not often encountered in a guideline. Guidelines are typically developed for topics for which there is controversy. Compelling inferences from first principles are usually not controversial and thus often are not selected to be the topic of a guideline.

Although rarely needed in a guideline, the use of compelling inferences from first principles without evidence is illustrated by the AAN's practice recommendations regarding the determination of brain death. That guideline determined that the evidence supporting the selection of a specific observation time to ensure irreversibility of the cessation of brain function was weak. Because of this, strong recommendations for choosing specific observation times before the declaration of brain death could not be made. Despite the absence of evidence, however, a compelling inference from first principles—in this case the requirement of irreversibility within the definition of brain death itself—supported a strong recommendation that clinicians must choose some observation period before the declaration of brain death to ensure that brain function did not return. In the guideline the selection of the specific duration of the observation period was left to physician judgment. Similar compelling inferences led to strong recommendations, despite the absence of evidence, that the clinician must know the proximate cause of the brain insult and must exclude confounding circumstances before declaring brain death.

A compelling inference from first principles alone is one circumstance in an AAN guideline where a strong recommendation could be developed in the absence of strong evidence.

Identifying Other Factors Affecting the Recommendation That Potentially Change the Recommendation Level

Although compelling inferences from first principles constitute one circumstance where a strong recommendation can be developed in the absence of strong evidence, there are other circumstances where non–evidence-based considerations will affect the strength of the recommendation.

Generalizability

As discussed in the section on formulation of conclusions, at times the evidence has limited generalizability. For example, the efficacy of steroids in patients with Bell's palsy may have been demonstrated by high-quality studies in women only. In this situation the recommendation could be crafted in a way that it is stronger when applied to women than when applied to men. Below is an example:

> Clinicians must offer steroids to women with new-onset Bell's palsy to increase the likelihood of facial function recovery (Level A). Clinicians should offer steroids to men with new-onset Bell's palsy to increase the likelihood of facial function recovery (Level B).

There is of course some judgment involved in deciding how generalizable the evidence is. The guideline developers should consider explicitly the issue of generalizability, come to a consensus, and transparently indicate their rationale. Reasonable guideline developers faced with the hypothetical gender-limited nature of the Bell's palsy evidence described above might alternatively conclude that there is no plausible biologic reason to conclude that steroids would not also help men. After transparently stating this assumption, they might craft a recommendation that reads as follows:

> Clinicians must offer steroids to patients with new-onset Bell's palsy to increase the likelihood of facial function recovery (Level A).

Clinical Importance of the Effect

At times high-quality evidence demonstrates an effect of therapy that is of marginal importance. For example, several Class I comparative trials of antiplatelet medications might show a statistically significant advantage of one drug over another. However, the effect is small. For every 150 patients treated with drug A instead of drug B over 2 years, only one stroke is prevented. This is a small effect. Even though the quality of evidence

is high, a lower level of recommendation seems justified. Thus, instead of stating drug A "must" be offered over drug B, it would be appropriate to state drug A "should" (or even "may") be offered over drug B.

The Risk of Harm of the Intervention and the Relative Value of the Benefit as Compared with the Risk

Harm includes issues of both tolerability (an unpleasant side effect that is not dangerous) and safety (a potentially dangerous side effect). Considerations of harm are dominated by issues of safety. Sometimes the evidence that is formally reviewed illuminates the frequency and magnitude of the potential harms of an intervention. When the harms are important (frequent or dangerous) it is most often useful to highlight them in the wording of the recommendation itself. For example, high-quality evidence indicates that a drug for secondary progressive MS dramatically reduces the risk of subsequent attacks (number of attacks reduced from an average of 2.3 a year to 0.6 a year) and cumulative disability but rarely (risk 1 in 1000) causes progressive multifocal encephalopathy (PML), a usually fatal condition. In this situation it is important that the recommendations describe both the benefit and harm.

For patients with secondary progressive MS with an attack frequency of $\geq$ 1 per year despite treatment with other MS therapies, clinicians should offer drug A to reduce MS attack rates and cumulative disability. The clinician must inform the patient of the risk of PML (1 in 1000) when discussing the potential risks and benefits of treatment (Level A).

Note that in this example the evidence level is indicated after the sentence describing the benefit and harm. This indicates that the evidence for harm was formally reviewed and rated during guideline development.

There are other situations where the evidence of an intervention's risk of harm was not part of the formal evidence base. The principle of care "first do no harm" justifies the inclusion of a statement regarding these harms in the recommendation even when the evidence is weak or not formally reviewed. This often happens when a rare but dangerous side effect is discovered during postmarketing surveillance. The evidence of harm may be based on weak evidence such as a case series or even isolated case reports. It is still important to include these potential harms in the recommendation itself. Assuming that this situation now applies to our example of MS drug use and PML risk, the recommendation might read as follows:

> For patients with secondary progressive MS whose attack frequency is $\geq$ 1 per year despite treatment with other MS therapies, clinicians should offer drug A to reduce MS attack rates and cumulative disability (Level B). The clinician should inform the patient of several isolated case reports of PML (exact risk unknown) when discussing the potential risks and benefits of treatment.

Here the level of evidence is parenthetically included only in the first sentence regarding benefit and not in the sentence describing the potential safety concern. This indicates that evidence regarding harm was not formally assessed in the guideline.

Not only can authors modify a recommendation to ensure that harms are described, but they can also downgrade the recommendation strength when warranted by the relative balance of benefit and risk. In extreme circumstances where the benefit-to-risk ratio is too close to call, the recommendation can be downgraded to the point that no recommendation can be given.

There are sophisticated techniques designed to measure quantitatively the balance and tradeoffs of the risks and benefits of an intervention. These include decision analysis and cost-effectiveness analysis. Generally, such analyses are beyond the scope of a guideline.

Availability, Cost, and Alternative Interventions

Strong evidence might support the use of an intervention that is unavailable or exorbitantly expensive. At a minimum, such issues should be discussed in the clinical context section. There may be times, particularly when cost is astronomical, that the strength or wording of the recommendation should be modified to convey these issues.

Additionally, there may be alternative therapies available for the condition of interest. Ideally comparative efficacy studies would be available to allow recommendations pertinent to the relative merits of one drug over another. Often such comparative evidence is not available. Even worse, the alternative therapy might not have been studied at all.

For example, amitriptyline might be the only drug studied for the treatment of depression in PD. Assume there is one Class I study showing benefit. Because of amitriptyline's side effect profile, the potential for harm, and the availability of potentially effective alternative therapy, it would be appropriate to craft the recommendation thusly:

> Clinicians may prescribe amitriptyline for patients with PD and depression to reduce depressive symptoms (Level C). Before prescribing amitriptyline clinicians should assess both the patient's ability to tolerate potential anticholinergic side effects and the patient's risk of cardiac dysrhythmias. Additionally, patients should be informed of the availability of alternative antidepressant therapies that have not been studied in PD but that have potentially better safety and tolerability profiles.

Although the evidence could have supported a Level B recommendation, the safety concerns and availability of alternative therapies led to a downgrading of the strength of the recommendation to Level C. The decision to discuss the side effects and presence of alternative therapies within the recommendation itself (after the designation of evidence level) or within the clinical context section needs to be made on a case-by-case basis by the guideline developers on the basis of the potential impact of the issues. Regardless of the decisions made, the decision and its rationale should be transparently indicated in the methods section of the manuscript.

Synthesis of All Factors and Determination of a Recommendation Level

It is evident that numerous factors can influence the wording and rating of a practice recommendation. Keeping track of these varying factors and their relative importance can be difficult. To assist in this process the AAN uses a graphical tool called a "Clinical Contextual Profile" (see appendix 7). The rows indicate each of the factors to consider in developing recommendations. The columns are labeled to aid you in making your judgment regarding the magnitude or importance of that factor relative to development of the recommendation. The output of the tool is indicated by the top row—the recommendation level.

To use the tool, first indicate the overall confidence in the evidence (second row from the bottom: High, Moderate, Low, or Very Low) and the strength of any deductive inferences

(bottom row: Compelling, Convincing, Plausible, Not Plausible). The lower of these two factors anchors the level of recommendation. For example, compelling deductive inferences with moderate supporting evidence would be anchored to a Level B recommendation (simply follow the column from Moderate up to Level B in the top row).

TIP

In unusual circumstances a recommendation may be based on deductive inferences from first principles alone. In this circumstance, the level of recommendation is anchored to the strength of the deductive inference only.

Next rate the magnitude of the other factors highlighted in the other rows. Any of these factors can be used to downgrade the recommendation level. For example, if patient values relative to potential benefits or risks of the outcome are judged to be highly variable, it is reasonable to downgrade our confidence that adherence to the recommendation will improve outcomes (because what is desirable varies from patient to patient). Likewise, we would not be confident that attempted adherence to a recommendation to use an intervention that has limited availability would improve outcomes—it might be appropriate to downgrade a recommendation or an intervention with limited availability. Moreover, even though we may be highly confident in the evidence relative to a specific intervention, if the relative value of the benefit versus risk is low, it may be appropriate to downgrade the recommendation level.

Only one factor—the relative value of benefit versus risk—can be used to upgrade the recommendation level from that determined by the confidence in the evidence or the strength of deductive inferences. If the relative value of the benefit versus the risk is judged by the author panel to be large or moderate, a lower recommendation level can be upgraded one level. On the basis of the relative value, the recommendation level cannot be upgraded by more than one level and can never attain a Level A rating.

Crafting the Recommendations

AAN practice recommendations must be actionable. The most important part of a recommendation is the verb (action word) used to indicate the action that should be taken. A good verb choice for a recommendation is unambiguous and indicates a specific action that the clinician should perform. Essaihi et al[1] have compiled a list of 11 suggested action verbs for guideline recommendation statements. These are: *test, prescribe, perform, educate/counsel, dispose, monitor, refer/consult, prepare, document, advocate,* and *diagnose/conclude.* The actions advised in most guideline statements should correspond to one of these 11 general action types. One of these terms (or a variant thereof) should be included in every AAN guideline recommendation statement.

DID YOU KNOW?

The step of "making recommendations" in the CPG development process necessarily requires the judgments—or opinions—of the guideline developers. Relying on opinions has a high risk of introducing bias. To minimize this risk the AAN has instituted the following steps:

1. Enforce a rigorous conflict of interest policy for guideline developers.
2. Obtain consensus from guideline developers using a modified Delphi process. This process involves anonymous voting, facilitated discussions, group feedback, and statistical analysis of the responses. The technique minimizes biases that can be introduced by group dynamics (e.g., group reinforcing extreme opinions) or dominant personalities.
3. Transparently describe difference of opinion.

The AAN has chosen to implement three basic recommendation levels: Level A, Level B, and Level C. Each level corresponds to a helping verb that denotes the level of obligation of the recommendation. Level A is the strongest recommendation level and is denoted by the use of the helping verb *must. Must* recommendations are rare, as they are based on the high confidence in the evidence and require both a high magnitude of benefit and low risk. Level B corresponds to the helping verb *should. Should* recommendations tend to be more common, as the requirements are less stringent but still based on the evidence and benefit-risk profile. Finally, Level C corresponds to the helping verb *may. May* recommendations represent the lowest allowable recommendation level the AAN considers useful and accommodate the highest degree of practice variation.

The wording of the recommendation needs to be modified in those circumstances where the evidence indicates that the intervention is not effective or useful. For example, if multiple adequately powered Class I studies demonstrate that an intervention is not effective, the recommendation could read "should *not* prescribe." See appendix 6 for a more in-depth discussion of suggested wording for conclusions and recommendations.

DID YOU KNOW?

The word *consider* should not enter into an AAN guideline recommendation statement. Research has shown that the word *consider* is confusing to guideline users,[2] and it is also difficult to quantify whether a person has effectively "considered" an action.

[1]Essaihi A, Michel G, Shiffman RN. Comprehensive categorization of guideline recommendations: creating an action palette for implementers. Musen M, ed. Proc Amer Med Informatics Assoc 2003;Washington, DC:220–224.

[2]Codish S, Shiffman RN. A model of ambiguity and vagueness in clinical practice guideline recommendations. AMIA Annu Symp Proc 2005;146–150.

Table 2 (see page 21) is a tool for building recommendations using AAN suggested verbiage. A more detailed tool is presented in appendix 6.

Basing Recommendations on Surrogate Outcomes

As previously stated, authors are urged to avoid using studies where only surrogate outcomes are measured, as it is often difficult to know the relevance of such outcomes. There are situations, however, where the studies providing strong evidence relevant to a topic measure surrogate outcomes only.

For example, there is controversy regarding the comparative effectiveness of brand-name versus generic antiepileptic drugs (AEDs). The only strong evidence available compares changes in serum AED levels in patients switched from brand-name AEDs to generic AEDs. Serum AED levels are, of course, a surrogate outcome. It is unclear how well they correlate with clinically meaningful outcomes such as seizure control and AED-related side effects. In this situation the AAN guideline development process permits authors to draw conclusions and make recommendations but only in reference to the surrogate outcome.

The conclusions and recommendations cannot imply an effect on clinically relevant outcomes.

For example, assuming multiple Class I studies show the lack of pharmacologic equivalence (within some prespecified serum AED-level threshold) you might conclude the following:

> Different formulations (generic, different nongenerics) of AEDs are highly likely not to be pharmacologically equivalent (multiple Class I studies).

Note that the conclusion discusses only the surrogate outcome (consistent serum AED levels).

Crafting the recommendation becomes problematic:

> For patients with epilepsy, the same formulations of AEDs should be offered to maintain consistent serum levels of the AED (Level A).

This actionable recommendation seems inappropriate because the benefit of stable AED levels is not a clinically important outcome.

The link between consistent levels and meaningful outcomes (seizure control, side effects) should be explicitly considered. If a compelling inference for this link cannot be derived from principles or strong evidence, a conditional recommendation can be made:

> For patients with epilepsy, if consistent serum AED levels are likely to improve seizure control or decrease the risk of toxicity, the same formulation of the AED should be used (Level A).

Knowing When Not to Make a Recommendation

When there is insufficient evidence to support or refute the effectiveness (or usefulness) of an intervention, no recommendation can be made. In such circumstances to highlight the lack of evidence state the following:

> No recommendation can be made because of insufficient evidence (Level U).

If the available evidence is insufficient to justify any practice recommendations, a systematic review (rather than a guideline) still can be published. Highlighting the gaps in evidence in such circumstances becomes particularly important. In the absence of recommendations the document is relabeled from "evidence-based guideline" to "evidence report" to denote the absence of recommendations.

Even when there is high-quality evidence, a recommendation need not necessarily follow. For example, there may be major concerns of generalizability or clinical applicability within the evidence base that would call into question the usefulness of any associated recommendations. In these circumstances, a formal recommendation is not required. A placeholder within the document where the recommendation would normally appear still needs to be present. This placeholder section would briefly explain why a recommendation was not made. In most circumstances, the limitations of the evidence resulting in the absence of a recommendation would be explicated in the published guideline.

Making Suggestions for Future Research

Often after formally reviewing the evidence, the guideline developers are in a unique position to suggest future research to fill in the evidence gaps. The future research section of the guideline is important for identifying areas that were found deficient on the basis of the thorough, systematic literature analysis.

Table 2. Elements of Recommendations

Mandatory Elements	Suggested Verbiage		
When (in what circumstances and in what patient population)	(For/In) patients with condition X		
Who (the person performing the action of the recommendation statement)	Clinicians		
Level of obligation (A, B, C)	A: Must (not) prescribe, offer (Rx) Must (not) test, counsel, monitor (Scrn, Dx, Px) Must avoid (causation)	B: Should (not) offer, prescribe Should (not) test, counsel, monitor Should avoid	C: May offer, prescribe May test, counsel, monitor, educate* May avoid May choose not to offer, prescribe May choose not to test, counsel, monitor
What (do what): Intervention (co-intervention): Intervention A (as compared with intervention B)	Describe specific intervention/test		
To precipitate what: (outcome)	Outcome Y		
Level of evidence: (Level N)			

*In the special case of negative Level C recommendations, we add the word *choose* because the term *may not* connotes a higher level of obligation than is intended.

Please see appendix 6 for additional guidance for constructing recommendations.

Logistics of the AAN Guideline Development Process

This section describes the logistics of AAN guideline development. It encompasses such topics as how to propose a guideline topic, how to conduct a literature search, and how to format and write an AAN guideline for publication.

Distinguishing Types of AAN Evidence-based Documents

The AAN develops systematic reviews and evidence-based guidelines to assist its members in clinical decision making—particularly in situations of controversy or variation in practice.

The AAN processes for developing systematic reviews and evidence-based guidelines are overseen by the GDS. The GDS reports to the AAN Practice Committee, and GDS members are appointed to 2-year terms by the AAN president. GDS members have expertise in systematic review methodology, guideline methodology, and representative subspecialties within neurology.

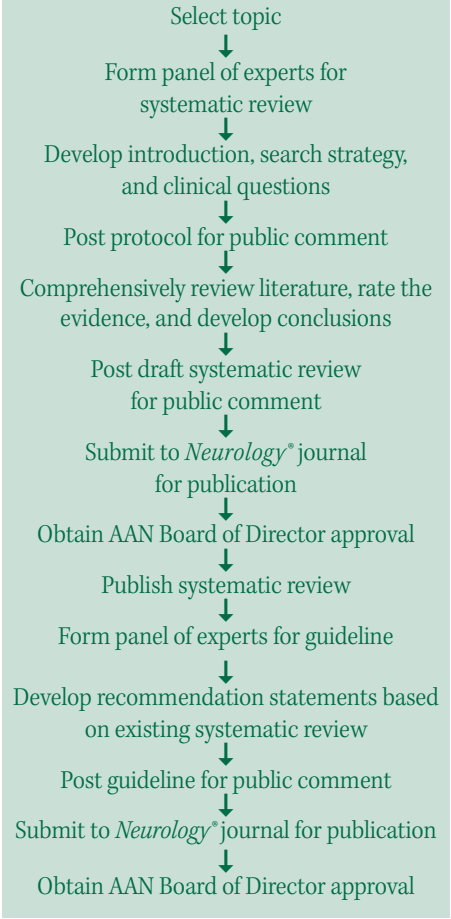
Expert author panels are formed for each project under development, to critically assess all relevant literature on a given topic or technology. Evidence is rated on the basis of quality of study design (systematic review), and clinical practice recommendations are developed and stratified to reflect the quality and content of the evidence. Evidence-based guidelines developed by the GDS are written with a patient-centric focus or an intervention-centric focus. Figure 7 depicts the steps of the AAN guideline development process.

The following are the key audiences of an AAN guideline:

Primary: Neurologists

Secondary: Patients, patient advocacy organizations, payers, federal agencies, (e.g., Centers for Medicare and Medicaid Services), clinical researchers, other health care providers

***Figure 7.* Steps in AAN Guideline Development**



Identifying the Three Document Types

Evidence Reports (Systematic Reviews)

These are documents developed using the AAN's EBM approach to developing guidelines. These documents do not include practice recommendations because of the lack of available high-quality published data. However, the reports provide neurologists with information about the state of the evidence and often serve as an impetus for researchers to design studies to address the current knowledge gaps.

Evidence-based Guidelines

These document types make actionable practice recommendations based on systematic reviews. As with evidence reports, guidelines are documents that assess the safety, utility, and effectiveness of new, emerging, or established therapies and technologies in the field of neurology. Contrary to evidence reports, evidence-based guidelines also address strategies for patient management that assist physicians and patients in clinical decision making, focusing on a series of specific, evidence-based practice recommendations that answer an important clinical question.

Case Definitions

Case definitions are documents developed for conditions for which there is no validated reference standard. In these circumstances, evidence cannot adequately define the condition; therefore these documents are developed using a formal, validated expert consensus approach (e.g., modified Delphi).

Understanding Common Uses of AAN Systematic Reviews and Guidelines

AAN systematic reviews and guidelines have the following uses:

- Improve health outcomes for patients
- Stay abreast of the latest in clinical research
- Provide medico-legal protection
- Advocate fair reimbursement
- Determine whether one's practice follows current, best evidence
- Reduce practice variation
- Affirm the role of neurologists in the diagnosis and treatment of neurologic disorders
- Influence public or hospital policy
- Promote efficient use of resources
- Identify research priorities on the basis of gaps in current literature

Nominating the Topic

Any AAN member, Committee, or Section, or an outside organization (e.g., an organization responsible for generating health policy) may submit a guideline topic nomination proposal. All topic nominations must be submitted in the form of a justification statement on *www.aan.com/guidelines.*

Periodically, the AAN Board of Directors will select a broad topic for the development of a set of practice guidelines. Broad topics targeted for guideline development in the past have included muscular dystrophies, PD, MS, epilepsy, dementia, and headache.

Topics are evaluated quarterly by the GDS Topic Review Panel, made up of a subset of the GDS members, which determines whether the topic is best addressed from the perspective of patient care or use of a technology or therapy.

The GDS officially approves a topic on the basis of neurologists' need for guidance, the availability of evidence to provide guidance, the potential to improve patient care and outcomes, and the availability of staff and resources.

Collaborating with Other Societies

After a topic is approved, the GDS may decide that the project would benefit from the perspective of other, related medical specialty societies. Obtaining this perspective is accomplished in the following ways:

- **Full collaboration:** The author panel reflects equal representation from the collaborating societies. The societies involved sign a formal letter of agreement outlining terms of copyright ownership, simultaneous publication, and division of costs prior to project initiation.
- **Invited participation:** AAN staff will work with the societies to obtain a nominee. This individual will act as the official representative from the organization, providing updates to the organization's board of directors. The organization will have an opportunity to review and comment on the systematic review and guideline during the public comment period and have an opportunity to endorse the systematic review and guideline prior to publication.

Forming the Author Panel (Bias/Conflict of Interest)

The GDS assigns a committee member to serve on a project facilitation team. A facilitation team member is designated the lead facilitator. The facilitation team helps with reviewing and rating the articles and grading the evidence. The facilitation team acts as the liaison to the GDS. In rare cases, the lead author of the systematic review will be the person who submitted the topic. The facilitation team, with the help of AAN staff and the GDS, assemble an author panel, being careful to seek a variety of perspectives, to avoid bias, and to avoid financial and intellectual conflicts. The author panel should always include a patient (when practical) and a patient advocate.

The panel should be capable of defining the clinical question(s) and performing the technical aspects of the systematic review development. It should be multidisciplinary in composition, with experts in systematic review methodology, including risk of bias assessment, study design, and data analysis; librarians or information specialists trained in searching bibliographic databases for systematic reviews; and clinical content experts to validate the questions and the search results. Clinical content experts will not review and rate the evidence. Other relevant users and stakeholders should be included as feasible. A single member of the review team can have multiple areas of expertise.[1,2]

The panel size (includes the facilitation team) will depend on the number and complexity of the question(s) being addressed. The author panel usually numbers between 5 and 10 individuals. The number of individuals with a particular expertise needs to be carefully balanced so that one group of experts is not overly influential.

Often, it is useful to have nationally recognized experts who are familiar with the literature pertaining to the topic being addressed (i.e., have authored clinical publications in high-impact journals). Participants with these credentials (rightly or wrongly) increase the credibility of the publication.

Revealing Conflicts of Interest

The AAN is committed to producing independent, critical, and truthful evidence-based guidelines. The AAN believes that those who produce the guidelines and those who have a financial stake in the success or failure of the products appraised in the guidelines should be kept separate and distinct. However, it may be difficult to form an expert panel devoid of potential conflicts of interest (COI). (see *http://www.aan.com/globals/axon/assets/3969.pdf*).

The COI policy of the AAN is strictly enforced for those members of the panel who participate in article selection, rating, and data extraction. When developing a systematic review, the AAN forms a panel free of intellectual, academic, and financial conflicts.

PITFALL

Special care should be taken when inviting nationally recognized experts in the field to serve as guideline authors. These authors may have predetermined ideas about the state of the evidence and with the structure of the conclusions and recommendations, and may not agree with the AAN's strict study-grading criteria. Additionally, they may have authored many of the studies of interest and be unable to provide an unbiased perspective.

[1]Institute of Medicine of the National Academies. Clinical Practice Guidelines We Can Trust: Standards for Developing Trustworthy Clinical Practice Guidelines (CPGs). *http://www.iom.edu/Reports/2011/Clinical-Practice-Guidelines-We-Can-Trust.aspx*. Released March 23, 2011. Accessed August 11, 2011.

[2]Institute of Medicine of the National Academies. Finding What Works in Health Care: Standards for Systematic Reviews. *http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx*. Released March 23, 2011. Accessed August 11, 2011.

Although individuals with a COI can be part of the CPG panel as a whole, panel members with financial or other important COIs cannot participate in these critical stages of the systematic review. The AAN carefully balances the panel composition between those with COIs (financial, research, academic, etc.) and those without when developing the guideline. Over half of the panel members should lack a COI. The chair of the panel must be free of any conflicts and remain conflict free for one year beyond the publication of the guideline. For guidelines of broad scope, panel members should not all be affiliated with the same institution or study group. If there is a recognized, credible controversy regarding the chosen guideline topic, both perspectives should be represented on the panel. In addition, authors of systematic reviews and CPGs are prohibited from serving as industry speakers or as expert witnesses for one year postpublication.

Obtaining Conflict of Interest Disclosures

Panel members must complete and sign a COI statement (see appendix 8) annually and as relationships change. All potential conflicts for the author, spouse, and minor children for the 2 years prior to filing the form must be disclosed. This form must be reviewed by the facilitation team, the AAN EBM methodologist, AAN staff, and GDS leadership before the prospective author panel member may commence work on the project.

Identifying Conflicts That Limit Participation

The GDS reserves the right to make changes to the author panel of the systematic review and the guideline to ensure balance and avoid bias. The GDS may choose not to appoint an individual as a lead author or as lead of a section of a guideline if the individual has any of the following relationships to the issues or products being assessed: having stock or stock ownership, being compensated for expert testimony, being a pioneer or having any substantial direct or indirect compensation or other relationship that GDS deems as creating a conflict. The lead author cannot have any financial or other important COI related to the guideline topic.

The AAN forbids commercial participation in guideline projects. Being a current employee of a pharmaceutical company or a device manufacturer precludes participation.

Disclosing Potential Conflicts of Interest

All disclosures will be published in the guideline as required by the *Neurology*® journal. In addition, a COI statement summarizing this policy will be included in all guidelines, as shown below.

Conflict of Interest Statement

The American Academy of Neurology is committed to producing independent, critical, and truthful systematic reviews and evidence-based guidelines. Significant efforts are made to minimize the potential for conflicts of interest to influence the recommendations of this systematic review and guideline. To the extent possible, the AAN keeps separate those who have a financial stake in the success or failure of the products appraised in the systematic reviews and guidelines and the developers of these same documents. Conflict of interest forms were obtained from all authors and reviewed by an oversight committee prior to project initiation. The AAN limits the participation of authors with substantial conflicts of interest. The AAN forbids commercial participation in, or funding of, systematic review and guideline projects. Drafts of the systematic review and guideline have been reviewed by at least three AAN committees, a network of neurologists, *Neurology*® peer reviewers, representatives from related fields, and the public. The AAN Guideline Author Conflict of Interest Policy can be viewed at *www.aan.com*.

Undertaking Authorship

All participating panel members, including the facilitator, are listed as authors. The lead author and facilitator determine the order of authorship and arbitrate any questions regarding who qualifies for authorship. The journal has strict guidelines regarding who should and should not be considered an author on the paper. At the time of submission to the journal, all author panel members will be required to complete a form affirming their contribution to the project as involving either study design/conceptualization, data/statistical analyses, or writing/revising of the manuscript. Authors whose work does not fit within any of these categories may not be authors but may be acknowledged as contributors in the acknowledgments section of the manuscript.

DID YOU KNOW?

All AAN systematic review and guideline authors perform the work of guideline authorship on behalf of the AAN. Therefore, the AAN is the sole owner of the rights to the guideline. Authors are required to transfer copyright to the AAN before work begins.

Understanding Roles and Responsibilities

Lead Facilitator

A GDS member is assigned to guide the project and advises on process issues—particularly the classification of evidence and translation of evidence to practice recommendations. This person reports to the GDS quarterly on project progress and may serve as the lead author of the systematic review and guideline.

Facilitation Team

A GDS member is assigned to guide the project. This person advises on process issues—particularly the classification of evidence and translation of evidence to practice recommendations.

Lead Author (If Different from the Lead Facilitator)

This person works with the lead facilitator to set timeline, assign tasks to panel members, and coordinate activities (e.g., literature review and drafting of the systematic review and guideline).

Author Panel Member

This person is an active participant in the project who usually reviews articles, classifies evidence, and writes portions of the systematic review and guideline.

EBM Methodologist

This person provides methodological and statistical guidance throughout the project, including assisting in forming clinical questions, developing data extraction forms, training authors on the AAN classification of evidence schemes, and adjudicating discrepancies in the rating of articles.

AAN Staff

AAN staff members provide administrative support and advice, facilitate meetings and group communications, provide manuscript management and copyediting (including for styles and standards), coordinate resource allocation (e.g., medical librarian), and coordinate the journal approval and publication process.

Completing the Project Development Plan

A project development plan (PDP) or project protocol outline is provided in appendix 9.

The following information is presented in the completed PDP:

- Justification for guideline
- Analytic frame used to help frame the questions
- Clinical questions (use the PICO format – reference section 2 of this manual)
- Terms and databases to be used in the literature search
- Inclusion and exclusion criteria for article selection
- Project timeline

Many of the elements of the PDP will inform the introduction and the section on description of the process in the final (published) guideline.

A PDP draft is submitted to the GDS and also is made publicly available on the AAN website for comment. After input from multiple reviews is received, the PDP is modified and finalized. A table of comments regarding the peer review and a list of corresponding changes (or the reasoning for changes not made) is developed by the guideline panel.

At times, the PDP will need to be revised during the course of guideline development. For example, exigencies may arise regarding

important additional questions that were identified throughout the course of the review; additionally, the authors may determine that the evidence available does not inform the questions originally developed but does inform a closely related question. Such changes to the PDP must be made with caution because bias may be introduced. Such changes can be made if deemed essential; the changes also must be documented in an amendment to the PDP.

Developing Clinical Questions

The AAN seeks focused, answerable clinical questions for systematic reviews and guidelines. A focused question makes the project more manageable and leads to conclusions and recommendations that are more pertinent to clinical care. Authors should select questions that can be answered on the basis of published, peer-reviewed evidence but also realize that AAN staff will make every effort to identify additional, relevant data in the gray literature.

The clinical question should address characteristics of the patients and interventions that are believed to significantly affect outcome. Taking too narrow a focus may unnecessarily limit the amount of evidence available for review. Conversely, taking too broad a focus or asking too many questions risks overwhelming the author panel with too much evidence and can encumber the process.

Remember, evidence-based guidelines are not textbooks or comprehensive summaries about how to diagnose and manage particular diseases. Rather, they are analyses of the published literature pertinent to specific aspects of care.

TIP

It may be helpful to perform a preliminary literature search to determine the availability of evidence to answer the questions being considered and to become familiar with the breadth of literature available on the topic.

Selecting the Search Terms and Databases

Search Terms

The authors should preliminarily identify the search terms that will ensure articles are obtained that can best answer the clinical questions. Authors should be sure to include appropriate synonyms from other nationalities, ethnicities, and disciplines.

Databases

The authors should identify the databases to be searched.

A MEDLINE search will likely uncover only 30% to 80% of published RCTs on a topic.[1] Therefore, it is recommended that authors search MEDLINE, EMBASE, and Science Citation Index or Current Contents. In consultation with a professional medical librarian, the author panel should determine on the basis of the topic being investigated whether it is appropriate to search additional databases. Some databases to consider are Bioethicsline, Cumulative Index to Nursing and Allied Health Literature (CINAHL), International Pharmaceutical Abstracts (IPA), Health Services Technology Assessment Texts (HSTAT), Psychological Abstracts, and BIOSIS. No guideline search should be limited to only one database; a minimum of two databases searched is required in the AAN process.

TIP

Authors should not be concerned with identifying all search terms at this stage. During the literature search process, the contracted medical librarian will suggest refinements and seek clarification of terms to ensure that the most comprehensive search is performed. It is essential for the content experts on the panel to identify a few key, relevant articles to ensure that these are identified by the search.

Selecting Inclusion and Exclusion Criteria

The author panel should develop criteria for including or excluding articles during the literature search and article review processes.

The criteria must be developed prior to beginning the search. However, they may be revised as necessary (e.g., if too few or too many studies are identified) as literature search results are obtained, provided that care is taken to avoid making changes that would introduce bias.

The author panel should develop an explicit list of inclusion and exclusion criteria by evaluating each of the following issues and any other issues that are pertinent to the specific topic being addressed. The GDS facilitation team can provide valuable assistance in completing this portion of the PDP.

[1]Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from *www.cochrane-handbook.org.*

Languages

Authors are encouraged to include all languages in the search, unless there is a specific reason for excluding non–English-language articles. English-language abstracts are available for many non–English-language articles. It is usually possible to obtain a translation of a non–English-language paper through a university, the Internet, or AAN staff.

Relevance

The type of subjects, intervention, and outcomes must be relevant to the clinical question.

Type of Subjects

Usually, the search is limited to articles concerned with human subjects. However, for some topics, it may be appropriate to include experimental articles from the laboratory. However, keep in mind that animal research does not drive conclusions or recommendations.

Intervention

The type of intervention should be explicit.

Outcome Measures

Outcome measures that will be examined should be included. Authors should consider whether the timing of follow-up for the outcome should be specified.

Types of Studies

The types of studies to be included in the search should be stipulated. If there is a large literature base, it may be appropriate to limit the search to RCTs (Class I) and controlled clinical trials (Class II). If the literature base is small, case control studies—and possibly retrospective case series—may be included. Authors should use methodological selection criteria only if doing so will result in obtaining articles that are clearly superior.

Setting the Project Timeline

A worksheet is provided on the PDP to help structure the project timeline. AAN staff members use the dates provided to develop an official project timeline that takes into account upcoming committee meeting dates and the availability of resources.

Performing the Literature Search

After the PDP is approved, the literature search should be conducted.

Consulting a Research Librarian

A medical librarian contracted by the AAN will develop and perform a comprehensive literature search based on the information given in the PDP regarding search terms and databases. The librarian will interactively query the database to define and refine the search as necessary. The lead guideline author and facilitator perform a quick review of the results on the basis of a preliminary search strategy, to ensure that key articles thought to be pertinent to the search are identified. When the search strategy is finalized through this iterative process, the strategy is sent to an independent research librarian for peer review.

Upon approval from the lead author and facilitator, the librarian completes the search of each database for each question. All results are compiled in an Endnote library and sent to AAN staff.

The content is uploaded into the AAN abstract database for the author panel's review.

Documenting the Literature Search

The literature search results are kept on file at the AAN. The following data are captured:

- Date(s) searches were conducted
- Search terms/strategy used
- Database(s) searched
- Date ranges included in search
- Explicit description of the inclusion and exclusion criteria

Documenting this information ensures the methods presented in the manuscript are transparent and reproducible. The entire search strategy for each question is published in the *Neurology*® journal as an appendix accompanying each systematic review and guideline.

Ensuring the Completeness of the Literature Search: Identifying Additional Articles

Upon receipt of the search results, the lead author and facilitator should critically evaluate the completeness of the search.

Authors should

- Ensure no essential concepts related to the question were missed
- Ask panel members to identify additional relevant articles (published or in press)
- Identify additional articles from reference lists, particularly the reference lists of review articles and meta-analyses

Using Data from Existing Traditional Reviews, Systematic Reviews, and Meta-analyses

Review articles can be categorized as traditional reviews, systematic reviews, and meta-analyses. Traditional reviews include publications such as book chapters, editorials, and expert reviews. Systematic reviews follow a rigorous methodology to address focused questions, apply explicit eligibility criteria, conduct exhaustive literature searches, and critically appraise the evidence. Meta-analyses consist of a systematic review plus statistical pooling of the results into a single summary measure, such as an odds ratio, relative risk, or risk difference. In addition, systematic reviews or meta-analyses may be embedded in such studies as economic evaluations, decision analyses, and CPGs.

Systematic reviews and meta-analyses are of particular importance in the AAN systematic review and guideline development processes. These studies contain much of the elements required for an evidence-based guideline (e.g., literature search, study selection, critical appraisal, and summary of results). Therefore, it is tempting to accept the study results at face value. However, there are several important disadvantages to this approach. Often small but important differences can be identified in the specific question addressed, the literature search, the definitions of clinical conditions and interventions, the thresholds for assessing outcomes, and the dates of the literature review. Furthermore, the evidence-rating systems of other organizations usually differ from the AAN's rating systems, and studies may not be described in sufficient detail to be rated according to AAN classification of evidence criteria.

Because of these disadvantages, usually traditional reviews, systematic reviews, and meta-analyses discovered during the literature search process will be used as follows:

A. Systematic reviews on the topic can be acknowledged in the clinical context section of the manuscript. This is encouraged when well-known systematic reviews have conclusions that contradict the guideline conclusions and recommendations.

B. The references cited in systematic reviews should be independently assessed for eligibility and then critically appraised and graded. (The reference list of the selected systematic reviews is compared with the results received from the literature review. Discrepancies are identified by AAN staff and sent to the lead author for input.)

C. Results and summary results (meta-analyses) of systematic reviews should not be used in drafting recommendations.

D. Results of individual studies as described within published systematic reviews should not be used at face value in drafting recommendations.

E. Differences in the results obtained by existing systematic reviews and AAN evidence-based guidelines should be acknowledged and explained in the text of the document.

Though not usually the case, at times a systematic review previously published elsewhere may address the same specific questions of a planned AAN practice guideline, and the methodological quality of the review may be substantially equivalent to that followed for an AAN evidence report. In these circumstances, after receiving GDS approval, the author panel can use the published systematic review as the basis for an AAN practice guideline.

Minimizing Reporting Bias: Searching for Non–peer-reviewed Literature

Often it is tempting to exclude non–peer-reviewed sources of evidence such as supplements, book chapters, and studies that are unpublished or are not included in bibliographic retrieval systems (so-called gray literature). Substantial empirical evidence demonstrates that excluding such evidence sources introduces bias. One major reason for this is that negative studies (i.e., studies not showing an effect of an intervention) are less likely to be published in peer-reviewed journals. These non–peer-reviewed sources provide important evidence that is not available in the peer-reviewed sources. Thus every effort should be made to assess this evidence to determine whether critical studies are being missed.

Reporting bias, including publication bias, presents a fundamental obstacle to the scientific integrity of systematic reviews and CPGs. To minimize the effect of reporting bias (and as recommended by the IOM) the AAN endorses a literature search process that

includes not only easy-to-access bibliographic databases but also other information sources that contain gray literature, particularly trial data and other unpublished reports. The search should be comprehensive and include both published and unpublished research. Additionally, systematic review and CPG panel members are encouraged to contact authors of primary studies under review, to clarify unclear reports or to obtain unpublished data that are relevant.

Selecting Articles

A two-step process is used to exclude articles that do not meet the inclusion criteria. All abstracts identified through the literature search are reviewed for relevance to the clinical question and adherence to the inclusion criteria. The same process is applied to the selected articles.

AAN staff will distribute the abstracts and articles and track panel member responses.

Reviewing Titles and Abstracts

Every abstract should be reviewed by at least two panel members. The lead author may choose to have two or more selected panel members review all abstracts or to have the abstracts distributed evenly among all panel members. When the number of authors reviewing the abstracts has been determined, AAN staff will use its guideline-reviewing database to assign the abstracts systematically to the authors, to ensure that each abstract is reviewed by two separate individuals.

Panel members review the abstracts and determine which are pertinent to the clinical question and meet the inclusion criteria. It is best to be inclusive at this stage of the process. If it is unclear whether an article is relevant or meets the inclusion criteria, it should be obtained for full-text review. If either reviewer indicates that an abstract is relevant, the associated article will be included for the full-text review. AAN staff will document the number of abstracts reviewed, the number excluded, and the reason(s) for exclusion.

Tracking the Article Selection Process

To ensure transparency of the systematic review and CPG development process, the disposition of every article identified by the search strategy should be tracked. The tracking should explicitly identify the reason for the exclusion of studies. A database of excluded study citations and the reason(s) for exclusion will be maintained online by AAN staff. After article selection, a flow diagram depicting the disposition of articles will be constructed (see above).

Obtaining and Reviewing Articles

After all abstracts are reviewed, AAN staff works with the authors to obtain and distribute the selected articles. Each article should be read independently by two panel members. The panel chair may choose to distribute the articles at random, by topic, or by another method.

NOTE

Because the AAN is not an educational institution, it is subject to copyright law. Thus the AAN cannot be treated as an interlibrary lender and must pay a copyright clearance center fee for each article selected by the authors that is not available free of charge in the public domain (average $35/article). Because AAN guideline searches are thorough and expansive, it becomes cost-prohibitive for the AAN to purchase each article. Therefore, AAN staff compiles a list of citations selected by the authors during abstract review, divides them according to the preference of the lead author and the author panel, and requests that authors use their institutional libraries to obtain personal copies of the article. In the event that costs are associated with an institutional library or research assistant to obtain the articles, the AAN will work with the institution to negotiate a fair price in order to ensure the guideline author is not incurring any costs.

Panel members should review each article for pertinence to the clinical question and adherence to the inclusion criteria set forth in the PDP. This is a screening review of the article; data should not be extracted at this point. It is best to be exclusive at this stage in the process. If it is unclear whether an article meets the inclusion criteria, it is appropriate to seek clarification through discussion with other panel members or by contacting the author of the study. However, if you choose to do the latter, you must contact all authors of studies for which you have similar questions, to avoid introducing bias.

If the panel members cannot agree on inclusion of a study, the study should be sent to an independent reviewer for adjudication. The adjudicator can be the lead author, any member of the facilitation team, a GDS member, or the AAN EBM methodologist.

Panel members send AAN staff a list of articles to be included in the guideline. AAN staff works with the lead author to compile a master list of articles to be included and resolves any disagreements regarding inclusion of individual articles.

Extracting Study Characteristics

The study characteristics—or elements—to be extracted from each article vary depending on the clinical question. In general, the characteristics extracted will correspond to one of the following categories:

- Citation information (depending on the software used, AAN staff will pre-populate this information)
- Items relevant to the study generalizability
- Elements relevant to the quality of evidence presented in the study
- Elements relevant to the study outcomes

TIP

The extraction of data and classification of evidence are crucial tasks. Many of the concepts discussed in this section are often unfamiliar to panel members who lack a methodological background. Panel members should seek the assistance of the facilitator in completing these steps, as necessary.

Developing a Data Extraction Form

The extraction of the study characteristics described above can be facilitated by development of a data extraction form. The AAN EBM methodologist, in conjunction with the lead author and the GDS lead facilitator, develops a data extraction form to apply to each clinical question. Sample data extraction forms are provided in appendix 10. It may be helpful for the facilitator or a GDS member to hold a conference call with all panel members to provide instruction prior to the start of data extraction.

Data from each article should be extracted by at least two panel members. Panel members complete these forms electronically, which are automatically submitted to AAN staff. Disagreement regarding the extracted elements, classification of evidence, or assessment of effect size should be resolved by consensus among panel members. If consensus

cannot be obtained, the GDS lead facilitator and a methodologist can arbitrate.

Constructing the Evidence Tables

Evidence tables are developed from the data extraction forms. The rows of the table correspond to each included study. The columns of the table correspond to the extracted study characteristics. It is essential to include the class of evidence determined for each study. Sample evidence tables can be found in appendix 5.

Sample table headings are provided below:

- Author, year
- Class of evidence (Class I, II, III, or IV)
- Purpose of study
- Study population: N, gender, mean age, diagnosis
- Interventions
- Outcome measures used
- Results (measure of association with a measure of statistical precision)

TIP

Tables are created in an electronic spreadsheet for easy manipulation. Evidence tables are required for each manuscript draft submission to the GDS.

Drafting the Document

The author panel should translate the evidence tables into a manuscript following the format provided in appendix 11.

Authors should use the following structural flow as described in the Preface:

clinical question → evidence → conclusions → recommendations*

*A recommendation section will be created only for a practice guideline.

Getting Ready to Write

Before authors begin writing the document, they should review appendix 11 in its entirety, as well as the "Instructions for Authors" and "Suggestions to Authors" at *www.neurology.org*. The manuscript will be evaluated by both the AAN guideline staff and *Neurology*® journal staff. It is essential to understand the expectations of each. The journal editorial policy limits the length of a print publication to a maximum of 3,500 words, 40 references published in the printed article, and 250 words in the abstract; however, authors of AAN systematic reviews and CPGs should focus primarily on adhering to the development requirements for these documents regardless of word count while being mindful of the need for succinctness in summarizing the evidence. AAN staff can help write the shorter version of the systematic review that will be published in print in the journal.

Usually, the lead author assigns specific topics to each author panel member; panel members develop the first draft of their assigned sections. The panel chair then integrates all of the sections into a cohesive document.

Formatting the Manuscript

The author panel should follow the structure provided in the manuscript format outlined here. AAN staff members with writing/editing expertise are available to assist in organizing the document, including populating standard text, numbering and formatting the references, and writing the abstract.

Drafts should be double-spaced and paginated, with text presented in Times New Roman 12-point font and line numbers included. Each draft should be labeled with the date and step in the process, as noted in appendix 11.

TIP

It is critical to be as transparent as possible in describing the process followed or results obtained in the development of the guideline. A long version of the final document will be published online, and a shorter version will be published in print.

Essential Elements

Cover Page

The cover page of the manuscript should include the title; author names, designations (MD, PhD, FAAN, etc.), and institutional affiliations; abstract word count; title character count (with spaces); manuscript word count; COI disclosures for all author panel members; and the date of draft. AAN staff can assist in writing COI disclosures, as staff keeps all forms on file.

Abstract

The abstract, although the first part of the manuscript presented, should be written last. The abstract is a brief, 250-word summary of the paper, highlighting the important points and findings. It is extremely difficult to write a 250-word summary of a manuscript not yet written. AAN staff is available to assist in drafting the abstract.

TIP

No information should be presented in the abstract that is not found in the manuscript, and all important points from the manuscript should be mentioned in the abstract. The abstract often is the only part of the article that physicians read, some of whom are reading it to determine whether to read the entire article.

The abstract should contain four sections: objective, methods, results (conclusions), recommendations. These are described as follows:

Objective: A brief, one-sentence statement regarding the purpose of the guideline (e.g., to perform an evidence-based review of the safety and efficacy of botulinum neurotoxin in the treatment of adult and childhood spasticity). The objective usually derives from the clinical questions.

Methods: A one- or two-sentence statement regarding the literature search strategy (including databases and years searched, if possible) and the method of classifying the evidence.

Results: Information from the conclusions sections of the paper.

Recommendations: A summary of the recommendations in the paper (along with their levels). Not all recommendations need to be presented. If there are many recommendations, it may be best to present only those with the strongest levels of evidence. Note that recommendations are not incorporated into the systematic review document.

Introduction

The introduction should be brief (no more than one or two pages). The introduction should include background on the topic (including prevalence, where applicable) and a brief description of gaps and controversies (i.e., a justification for this publication), and should end with a statement of the clinical questions to be examined in the rest of the manuscript. Explicitly state any assumed principles of care.

Description of the Analytic Process (Methods)

The description of the analytic process describes the exact process that the panel used to create the document. It is important that the description be detailed enough to be transparent and replicable. This section should describe panel formation (usually a

brief sentence stating that the AAN convened an expert panel made up of neurologists and [insert other specialists]), the literature review dates and databases searched, the secondary search strategy (usually examining the references of review articles), inclusion/exclusion criteria used, how articles were reviewed, the classification of evidence schemes used, the process by which authors resolved disagreements in classification, and any modifications to the schemes employed that were specific to this question. (Note that the complete search strategy will be presented in an online appendix.) This section should also describe the outcomes of interest, the measure of effect preferred, and the measure of statistical precision used, and should identify what was considered a clinically important effect.

Analysis of Evidence

This section is best organized by clinical question. Each clinical question is listed as a subheading under which the relevant evidence for that question is presented. Each subsection (clinical question) should provide the number of citations retrieved for that question at each stage of review: first abstracts, then full-text articles, and then the articles selected for incorporation into the paper. Describe the evidence, briefly justify the evidence rating, and provide appropriate quantitative measures of effect size, including measures of statistical precision (e.g., 95% confidence intervals) where possible.

Conclusions

A conclusions section should follow each clinical question subsection in the analysis of evidence section. Conclusions should be directly linked to the evidence and should use standard AAN conclusion language when possible (see appendix 6 for suggested language). For each conclusion, mention the number of supporting studies and the class of evidence and statistical precision of those studies. An example is, drug A is probably useful to reduce the symptoms of disease X (two adequately precise Class II studies).

Clinical Context

A description of the factors that influenced the recommendation should be summarized in a section preceding the recommendation. The primary purpose is to explain the rationale for the formulation of the specific recommendation. This section is labeled "Putting the evidence into a clinical context," or "Clinical Context."

This section may include any information that does not directly follow from the evidence presented. Such information includes alternatives for which there was limited evidence, risk-benefit profiles, limits to the generalizability of the evidence, magnitude of benefit, harms, cost, and outcomes not addressed in the evidence. This section can be presented after the recommendations section of each clinical question (as needed) or at the end of the manuscript.

Care must be exercised regarding the wording of this section so as to avoid the inclusion of any commentary that could be construed as recommendations based not on the evidence but rather on prevailing practice or opinion. To prevent this potential undermining of the careful, rigorous process used to develop AAN guidelines, the following process should be followed:

- First, consider whether the point to be made would be most appropriately addressed in the introduction rather than in a separate Clinical Context section.
- Leave the evidence-based recommendations unchanged.
- Include a description of the clinical context issue in paragraph form. Include critical issues only. No new recommendations can be made in the Clinical Context section.

If clinical context includes discussion of commonly used therapies or procedures excluded from the guideline because of lack of evidence, such therapies or procedures should be identified not as "standard of care" but rather simply as "common practice" and must include a relevant reference citation.

Recommendations (Included Only in Guidelines)

Recommendations are presented as a separate publication after all of the evidence for all questions has been presented. The recommendations should stem from the conclusions in the systematic review and should use standard AAN recommendation language when possible (see appendix 6 for suggested language). For each recommendation, a quality of evidence label (e.g., Level A) must be included. Recommendations should be written to support patient-centered outcomes* and should include a statement of harm (ideally a number needed to treat), if appropriate.**

*Avoid wording recommendations as such: "Therapy X should be prescribed by clinicians." Instead, restructure the wording as such: "Therapy X should be prescribed by clinicians to reduce spasticity in children and adolescents with cerebral palsy." The latter example presents a patient-centered outcome, which will aid clinicians in applying the guidelines in their practice.

**A recommendation should include a statement of harm especially when there are important or severe side effects, defined as those that may be life-threatening, are common and affect safety or quality of life, or are covered by a US Food and Drug Administration black box warning.

Recommendations for Future Research

The completion of the systematic review and analysis of the literature position the author panel favorably to recommend areas of future research. The future research section should present a summary of study design concerns that were found to be limitations in the existing literature, such as the need for multicenter studies, adequate sample sizes, randomized studies, and more comprehensive or reliable outcomes measures. This section should also address the need for more studies on therapies for which evidence was deemed inadequate or conflicting.

Disclaimer

This is a stock language statement provided by AAN staff.

Acknowledgments

The acknowledgments section is optional and is reserved for those who assisted in manuscript development but who do not qualify as authors under the *Neurology*® journal authorship policy. People who are frequently acknowledged are research assistants, editors, AAN staff, or reviewers who made significant comments.

DID YOU KNOW?

The *Neurology*® journal requirements for word count in AAN systematic reviews and guidelines include only the body of the paper, from the introduction through the recommendations for future research. Word count does not include the text of appendices, stock AAN language, references, or tables or figures.

Tables/Figures

Evidence tables are usually published online only as data supplements, although they are required to be presented as part of the manuscript at each stage of review. Figures may be published in print or online as necessary.

Appendices

The appendices include the GDS member roster, classification of evidence scheme, and classification of recommendations scheme. AAN staff provides this language, which is usually published online only.

References

Because of journal space requirements, the number of print references is limited to 40. The remaining references will be published online as e-references, the process for which AAN staff will coordinate.

Reviewing and Approving Guidelines

Stages of Review

AAN staff and the GDS will review the systematic review and CPG at several stages during the development process. These stages are outlined below:

Stage	Reviewers
General topic	GDS
Author panel* compositin	AAN staff and GDS leadership
Protocol	AAN staff, public, GDS
Evidence tables	AAN staff, GDS
Systematic review/ CPG draft GDS	AAN staff, peer review network, public
GDS-approved evidence report or practice guideline	Practice Committee, *Neurology*® peer eviewers, AAN Board of Directors

*The author panel includes members of the facilitation team.

These levels of review are described in more detail next.

AAN Staff-level Review

All initial draft documents, including the protocol, evidence tables, systematic review, and CPG, are first reviewed by AAN staff and the AAN EBM methodologist. These reviews ensure that drafts submitted to the GDS meet AAN requirements for methodological quality and formatting. Often, this step involves AAN staff queries to the author.

The most common revision requests pertain to the following:

- Poorly constructed clinical questions
- Incorrect classification of the evidence
- Significant deviations from the established format
- Incorrect translation of the evidence to conclusions
- Incorrect translation of the conclusions to recommendations
- Manuscript length (too long)

Author Response to Reviews

At each level of review, the author panel should revise the document, as appropriate, and populate a revision table that lists each reviewer, the reviewer comment, and how the comment was addressed in the document (see example in appendix 12). The revision table must be submitted to the GDS with each draft. The table will accompany the document when it is sent to the Practice Committee, the *Neurology*® journal, and the AAN Board of Directors.

Authors are encouraged to use electronic word-processing formatting tools (underline and strikethrough) or Track Changes for this draft and subsequent drafts for which the changes are minor.

The revised manuscript and revision table must be submitted to AAN staff.

Initial GDS review

After approval from AAN staff and the AAN EBM methodologist, draft documents are submitted to the GDS for review at one of its quarterly meetings. The GDS carefully reviews the documents and often requests revisions. AAN staff and the facilitator compile the list of requested revisions in a revision table (see appendix 12), and authors are asked to respond to all comments and revise the documents accordingly prior to the next GDS in-person meeting. The typical timeframe for revision of the manuscript is 6 to 8 weeks.

Public Comment

When the draft protocol, systematic review, or CPG receives initial GDS approval, AAN staff posts it for review and comment on *www.aan.com* for 30 days. The documents are shared publicly because the AAN realizes that systematic review and CPG development groups are limited to a small number of individuals for expediency and efficiency in the development process. The AAN will not limit the completion of the review to a predefined external reviewing group; rather, any individual will be able to access the document on *www.aan.com* for review and comment.

The AAN realizes that although the document may be publicly available at *www.aan.com* for 30 days, organizations may not be aware of its availability for comment. Thus, the AAN will invite the following individuals and groups to comment:

- AAN membership
- Members of AAN Sections
- Members of AAN Committees (including the Ethics, Law, and Humanities Committee), Subcommittees, Task Forces, and Work Groups

TIP

Your reference list should not be alphabetized but instead should be presented in the order referenced in the manuscript. You may use automatic reference numbering, such as Endnote Cite While You Write or Microsoft Word Endnotes; however, prior to submission to the journal all automatic reference numbering must be removed. Usually the AAN staff takes responsibility for fact checking references and assuring they are formatted correctly for *Neurology*® submission.

Use "people-first" language. For example, say "patients with dementia" rather than "demented patients."

The word data is plural (as in "data are," not "data is").

When referring to the class of a study, make sure to use Roman numerals (Class I, II, III) to avoid potential confusion from use of multiple numeric values, as in "2 Class 2 studies."

Always capitalize the word class when referring to a specific study (e.g., "Class I") and level when referring to a specific recommendation (e.g., "Level A").

- Domestic and international subject matter experts
- AAN legal counsel
- Physician organizations with a stake in the manuscript content (identified by the GDS, author panel, or AAN staff)
- Patient advocacy organizations with a stake in the manuscript content (identified by the GDS, author panel, or AAN staff)

Staff collects the responses and forwards them to the lead facilitator and lead author. The responses are presented in a revision table (see Appendix 11), and authors are required to respond to all reviewer comments. The authors decide whether to make changes to the manuscript on the basis of reviewer comments; however, the authors must adequately defend this decision in the revision table.

GDS Re-review (Post–public Comment)

AAN staff sends the GDS the revised documents and revision table (reflecting input from public comment) for a final review and an official vote at the next GDS meeting. GDS approval may be contingent on additional requested revisions.

Practice Committee Review

When the GDS gives final approval of the manuscript, AAN staff submits the manuscript to the Practice Committee for approval. The Practice Committee may have additional revision requests, and if these revisions are substantial, the changes are reviewed by the GDS chairs. Extremely substantial revisions—particularly those that change the conclusions and recommendations—may require GDS reapproval.

Journal Review

When the Practice Committee has approved the manuscript, AAN staff sends it to the *Neurology*® journal for peer review.

The *Neurology*® journal will solicit reviewers from its network to review and comment on the manuscript. Comments are sent directly to AAN staff. AAN staff will draft a revision letter on behalf of the lead author, presenting all comments from *Neurology*® peer reviewers. Authors are encouraged to consider all revisions suggested by the journal peer reviewers. Authors should contact the facilitator if the reviewers' requested changes conflict with AAN requirements for systematic reviews or CPGs, particularly if reviewers request substantial revisions to the wording of conclusions or recommendations.

The lead author should submit the revised draft to AAN staff (not directly to the journal) with the completed revision letter denoting the panel's responses to all of the journal reviewers' comments. The revised draft *must* show all changes made to the manuscript, using the electronic editing tool (e.g., Track Changes, strikethrough, or highlight). AAN staff then resubmits the manuscript to the journal.

At the time of revision submission, authors will be required to complete an authorship agreement form and a disclosure agreement form at the *Neurology*® website (*www.neurology.org*).

The journal may request additional rounds of reviews prior to accepting the manuscript for publication.

AAN Board of Directors Approval

When the manuscript has been accepted for publication in the *Neurology*® journal, AAN staff submits it to the AAN Board of Directors for approval.

Requests for revision during the approval process are reviewed by the GDS chairs. Substantive revisions may require reapproval by the GDS and Practice Committee.

Endorsement

It may be appropriate to seek guideline endorsement from other, relevant organizations. Authors should inform AAN staff of any organizations not identified during public comment from which to seek endorsement.

Taking Next Steps (Beyond Publication)

Undertaking Dissemination

At a minimum, the following steps are taken to promote an AAN systematic review and CPG:

- Published in *Neurology*® journal
- Posted on the AAN website
- Announced to AAN members by all-member email
- Announced in *AANnews*® and *AANe-news*
- Submitted to guideline compendia such as the National Guidelines Clearinghouse

The Practice Improvement Subcommittee, AAN guideline staff, or AAN public relations staff may undertake additional dissemination and implementation efforts. These may include development of a press release, slide presentation, clinical case example, summary of the systematic review and CPG for patients, summary of the systematic review and CPG for clinicians, and algorithms, to help members incorporate guideline recommendations in practice.

Responding to Correspondence

Because AAN staff members coordinate the journal submission and publication process, they receive any related letters to the editor. For any letters received, systematic review and CPG authors and facilitators should work together to draft a response letter. The response letter is reviewed internally by AAN staff and the subcommittee leadership prior to its submission to the journal.

LESSON LEARNED

Do not be discouraged if you receive a negative letter to the editor about your publication. These are typically submitted by other *Neurology*® journal authors who disagree with the guideline conclusions and recommendations. The AAN views such correspondence as opportunities to educate *Neurology*® journal readership on EBM principles.

Updating Systematic Reviews and CPGs

According to the Agency for Healthcare Research and Quality, guidelines have a 10% chance of being out-of-date 3 years from publication. Therefore the AAN has approved the system described below for evaluating systematic reviews and CPGs to ensure that out-of-date guidelines are identified and updated in a timely manner.

Biennial Review: Updating the Literature Search and Assessing Methodological Soundness

Systematic reviews and CPGs are assessed every 2 years to determine whether new literature has been published that would warrant an update. The following steps are taken:

- Biennial correspondence is sent to all authors and the facilitator.
- An updated literature search and a review of methodological soundness are performed by a GDS member. (Note: The search should specifically seek to identify new evidence that would change the conclusions in the systematic review or recommendations in the CPG.)

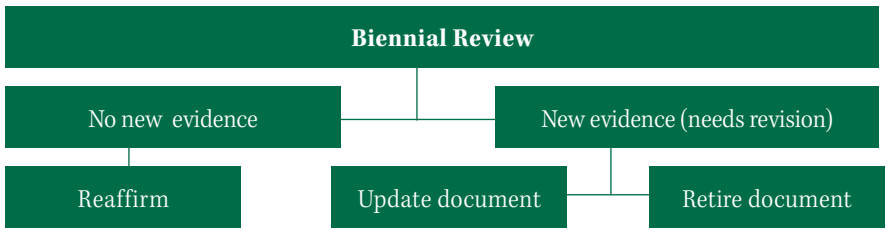
If an update is deemed warranted, GDS forms a new author panel, which may include members of the initial author panel. The project then follows the same process as outlined in this manual.

On occasion, the GDS will decide not to revise a document in need of updating. In these circumstances, the document will be retired.

Decisions regarding the update status will be communicated to the AAN membership through the AAN website. All documents biennially reviewed by the GDS that don't require an update are reaffirmed. Documents that require updating will be designated as such on *www.aan.com*. The status and date of the update action will be indicated on the website.

Figure 8 summarizes the AAN guideline update process.

Figure 8. **AAN Guideline Update Process**



Appendices

Appendix 1: Evidence-based Medicine Resources

Regarding Evidence-based Medicine and Reviews:

The Evidence-based Medicine Toolkit (available at *www.aan.com*)

Cochrane Handbook (available at *www.cochrane.org/training/cochrane-handbook*)

Evidence-Based Medicine (Sackett et al, 1997)

Evidence-Based Principles and Practice (McKibbon, 1999)

Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews (Academia and Clinic: Systematic Review Series). Ann Intern Med 1997;127:380–387.

National Guideline Clearinghouse (available at *www.guidelines.gov*)

The CATbank (available at *www.minervation.com/cebm2/docs/catbank.html*)

Regarding Using EndNote to Search Remote Databases:

www.biomed.lib.umn.edu/endref.html

Regarding Using EndNote to Create a Bibliography:

www.biomed.lib.umn.edu/end.html

Appendix 2: Formulas for Calculating Measures of Effect

Therapeutic Questions

	Good	Poor
Treated	A	C
Untreated	B	D

Relative rate = [A / (A + C)] / [B / (B + D)]
Rate difference = [A / (A + C)] - [B / (B + D)]

Diagnostic (Prognostic) Accuracy Questions

	Disease (Outcome) Present	Disease (Outcome) Absent
Test (Predictor) Positive	A	C
Test (Predictor) Negative	B	D

Relative risk = [A / (A + C)] / [B / (B + D)]
Sensitivity = A / (A + B)
Specificity = D / (C + D)
Positive predictive value = A / (A + C)
Negative predictive value = D / (B + D)

Screening Questions

	Condition Present	Condition Absent
Tested	A	C

Yield = A / (A + C)

Appendix 3: Classification of Evidence Matrices

Classification of Evidence Matrix for Therapeutic, Causation, and Prognostic Questions

	Clinical Question Type		
Class	**Therapeutic**	**Causation**	**Prognostic**
I	- Randomized, controlled clinical trial (RCT) in a representative population - Masked or objective outcome assessment - Relevant baseline characteristics are presented and substantially equivalent between treatment groups, or there is appropriate statistical adjustment for differences - Also required: a. Concealed allocation b. Primary outcome(s) clearly defined c. Exclusion/inclusion criteria clearly defined d. Adequate accounting for dropouts (with at least 80% of enrolled subjects completing the study) and crossovers with numbers sufficiently low to have minimal potential for bias e. For noninferiority or equivalence trials claiming to prove efficacy for one or both drugs, the following are also required*: 1. The authors explicitly state the clinically meaningful difference to be excluded by defining the threshold for equivalence or noninferiority 2. The standard treatment used in the study is substantially similar to that used in previous studies establishing efficacy of the standard treatment (e.g., for a drug, the mode of administration, dose, and dosage adjustments are similar to those previously shown to be effective) 3. The inclusion and exclusion criteria for patient selection and the outcomes of patients on the standard treatment are comparable to those of previous studies establishing efficacy of the standard treatment 4. The interpretation of the study results is based on a per-protocol analysis that accounts for dropouts or crossovers	- Cohort survey with prospective data collection - All relevant confounding characteristics are presented and substantially equivalent between comparison groups or there is appropriate statistical adjustment for differences - Outcome measurement is objective or determined without knowledge of risk factor status - Also required: a. Primary outcome(s) defined b. Exclusion/inclusion criteria defined c. Accounting of dropouts (with at least 80% of enrolled subjects completing the study)	- Cohort survey with prospective data collection - Includes a broad spectrum of persons at risk for developing the outcome - Outcome measurement is objective or determined without knowledge of risk factor status - Also required: a. Inclusion criteria defined b. At least 80% of enrolled subjects have both the risk factor and outcome measured
II	- Cohort study meeting criteria a–e (see Class I) or an RCT that lacks one or two criteria b–e (see Class I) - All relevant baseline characteristics are presented and substantially equivalent among treatment groups, or there is appropriate statistical adjustment for differences - Masked or objective outcome assessment	- Cohort study with retrospective data collection or case-control study. Study meets criteria a–c (see Class I) - All relevant confounding characteristics are presented and substantially equivalent among comparison groups or there is appropriate statistical adjustment for differences - Masked or objective outcome assessment	- Cohort study with retrospective data collection or case-control study. Study meets criteria a and b (see Class I) - Includes a broad spectrum of persons with and without the risk factor and the outcome - The presence of the risk factor and outcome are determined objectively or without knowledge of one another

Appendix 3: Classification of Evidence Matrices (Continued from page 35)

Classification of Evidence Matrix for Therapeutic, Causation, and Prognostic Questions			
Class	**Therapeutic**	**Causation**	**Prognostic**
III	- Controlled studies (including well-defined natural history controls or patients serving as their own controls) - A description of major confounding differences between treatment groups that could affect outcome** - Outcome assessment masked, objective, or performed by someone who is not a member of the treatment team	- Cohort or case-control study designs - A description of major confounding differences between risk groups that could affect outcome** - Outcome assessment masked, objective or performed by someone other than the investigator that measured the risk factor	- Cohort or case control study - Narrow spectrum of persons with or without the disease - The presence of the risk factor and outcome are determined objectively, without knowledge of the other or by different investigators
IV	- Did not include patients with the disease - Did not include patients receiving different interventions - Undefined or unaccepted interventions or outcome measures - No measures of effectiveness or statistical precision presented or calculable	- Did not include persons at risk for the disease - Did not include patients with and without the risk factor - Undefined or unaccepted measure of risk factor or outcomes - No measures of association or statistical precision presented or calculable	- Did not include persons at risk for the outcome - Did not include patients with and without the risk factor - Undefined or unaccepted measures of risk factor or outcomes - No measures of association or statistical precision presented or calculable

*Numbers 1–3 in Class Ie are required for Class II in equivalence trials. If any one of the three is missing, the class is automatically downgraded to Class III

**Objective outcome measurement: an outcome measure that is unlikely to be affected by an observer's (patient, treating physician, investigator) expectation or bias (e.g., blood tests, administrative outcome data)

Appendix 3: Classification of Evidence Matrices (Continued from page 36)

Classification of Evidence Matrix for Therapeutic, Causation, and Prognostic Questions		
	Clinical Question Type	
Class	**Diagnostic Accuracy**	**Population Screening**
I	- Cohort survey with prospective data collection - Includes a broad spectrum of persons suspected of having the disease - Disease status determination is objective or made without knowledge of diagnostic test result - Also required: a. Inclusion criteria defined b. At least 80% of enrolled subjects have both the diagnostic test and disease status measured	- Study of a cohort of patients at risk for the outcome from a defined geographic area (i.e., population based) - The outcome is objective - Also required: a. Inclusion criteria defined b. At least 80% of patients undergo the screening of interest
II	- Study of a cohort of patients at risk for the outcome from a defined geographic area (i.e., population based) - The outcome is objective - Also required: a. Inclusion criteria defined b. At least 80% of patients undergo the screening of interest	- A non–population-based nonclinical cohort (e.g., mailing list, volunteer panel) or a general medical, neurology clinic/center without a specialized interest in the outcome. Study meets criteria a and b (see Class I) - The outcome is objective
III	- Cohort or case control study - Narrow spectrum of persons with or without the disease - The diagnostic test result and disease status are determined objectively, without knowledge of the other or by different investigators	- A referral cohort from a center with a potential specialized interest in the outcome
IV	- Did not include persons suspected of the disease - Did not include patients with and without the disease - Undefined or unaccepted independent reference standard - No measures of diagnostic accuracy or statistical precision presented or calculable	- Did not include persons at risk for the outcome - Did not statistically sample patients or patients specifically selected for inclusion by outcome - Undefined or unaccepted screening procedure or outcome measure - No measure of frequency or statistical precision calculable

Appendix 4: Narrative Classification of Evidence Schemes

Therapeutic

Class I

- Randomized, controlled clinical trial (RCT) in a representative population
- Masked or objective outcome assessment
- Relevant baseline characteristics are presented and substantially equivalent between treatment groups, or there is appropriate statistical adjustment for differences
- Also required:
  a. Concealed allocation
  b. Primary outcome(s) clearly defined
  c. Exclusion/inclusion criteria clearly defined
  d. Adequate accounting for dropouts (with at least 80% of enrolled subjects completing the study) and crossovers with numbers sufficiently low to have minimal potential for bias
  e. For noninferiority or equivalence trials claiming to prove efficacy for one or both drugs, the following are also required*:
     1. The authors explicitly state the clinically meaningful difference to be excluded by defining the threshold for equivalence or noninferiority
     2. The standard treatment used in the study is substantially similar to that used in previous studies establishing efficacy of the standard treatment (e.g., for a drug, the mode of administration, dose, and dosage adjustments are similar to those previously shown to be effective)
     3. The inclusion and exclusion criteria for patient selection and the outcomes of patients on the standard treatment are comparable to those of previous studies establishing efficacy of the standard treatment
     4. The interpretation of the study results is based on a per-protocol analysis that accounts for dropouts or crossovers

Class II

- Cohort study meeting criteria a–e (see Class I) or an RCT that lacks one or two criteria b–e (see Class I)
- All relevant baseline characteristics are presented and substantially equivalent among treatment groups or there is appropriate statistical adjustment for differences
- Masked or objective outcome assessment

Class III

- Controlled studies (including well-defined natural history controls or patients serving as their own controls)
- A description of major confounding differences between treatment groups that could affect outcome**
- Outcome assessment masked, objective or performed by someone who is not a member of the treatment team.

Class IV

- Did not include patients with the disease
- Did not include patients receiving different interventions
- Undefined or unaccepted interventions or outcome measures
- No measures of effectiveness or statistical precision presented or calculable

*Numbers 1–3 in Class Ie are required for Class II in equivalence trials. If any one of the three is missing, the class is automatically downgraded to Class III

**Objective outcome measurement: an outcome measure that is unlikely to be affected by an observer's (patient, treating physician, investigator) expectation or bias (e.g., blood tests, administrative outcome data)

Causation

Class I

- Cohort survey with prospective data collection
- All relevant confounding characteristics are presented and substantially equivalent between comparison groups or there is appropriate statistical adjustment for differences
- Outcome measurement is objective or determined without knowledge of risk factor status
- Also required:
  a. Primary outcome(s) defined
  b. Exclusion/inclusion criteria defined
  c. Accounting of dropouts (with at least 80% of enrolled subjects completing the study)

Class II

- Cohort study with retrospective data collection or case-control study. Study meets criteria a–c (see Class I)
- All relevant confounding characteristics are presented and substantially equivalent among comparison groups or there is appropriate statistical adjustment for differences
- Masked or objective outcome assessment

Class III

- Cohort or case-control study designs
- A description of major confounding differences between risk groups that could affect outcome**
- Outcome assessment masked, objective or performed by someone other than the investigator that measured the risk factor

Class IV

- Did not include persons at risk for the disease
- Did not include patients with and without the risk factor
- Undefined or unaccepted measure of risk factor or outcomes
- No measures of association or statistical precision presented or calculable

**Objective outcome measurement: an outcome measure that is unlikely to be affected by an observer's (patient, treating physician, investigator) expectation or bias (e.g., blood tests, administrative outcome data)

Prognostic Accuracy

Class I

- Cohort survey with prospective data collection
- Includes a broad spectrum of persons at risk for developing the outcome
- Outcome measurement is objective or determined without knowledge of risk factor status
- Also required:
  a. Inclusion criteria defined
  b. At least 80% of enrolled subjects have both the risk factor and outcome measured

Class II

- Cohort study with retrospective data collection or case-control study. Study meets criteria a and b (see Class I)
- Includes a broad spectrum of persons with and without the risk factor and the outcome
- The presence of the risk factor and outcome are determined objectively or without knowledge of one another

Class III

- Cohort or case control study
- Narrow spectrum of persons with or without the disease
- The presence of the risk factor and outcome are determined objectively, without knowledge of the other or by different investigators

Class IV

- Did not include persons at risk for the outcome
- Did not include patients with and without the risk factor
- Undefined or unaccepted measures of risk factor or outcomes
- No measures of association or statistical precision presented or calculable

Diagnostic Accuracy

Class I

- Cohort survey with prospective data collection
- Includes a broad spectrum of persons suspected of having the disease
- Disease status determination is objective or made without knowledge of diagnostic test result
- Also required:
  a. Inclusion criteria defined
  b. At least 80% of enrolled subjects have both the diagnostic test and disease status measured

Class II

- Cohort study with retrospective data collection or case-control study. Study meets criteria a and b (see Class I)
- Includes a broad spectrum of persons with and without the disease
- The diagnostic test result and disease status are determined objectively or without knowledge of one another

Class III

- Cohort or case control study
- Narrow spectrum of persons with or without the disease
- The diagnostic test result and disease status are determined objectively, without knowledge of the other or by different investigators

Class IV

- Did not include persons suspected of the disease
- Did not include patients with and without the disease
- Undefined or unaccepted independent reference standard
- No measures of diagnostic accuracy or statistical precision presented or calculable

Population Screening

Class I

- Study of a cohort of patients at risk for the outcome from a defined geographic area (i.e., population based)
- The outcome is objective
- Also required:
  a. Inclusion criteria defined
  b. At least 80% of patients undergo the screening of interest

Class II

- A non–population-based nonclinical cohort (e.g., mailing list, volunteer panel) or a general medical, neurology clinic/center without a specialized interest in the outcome. Study meets criteria a and b (see Class I)
- The outcome is objective

Class III

- A referral cohort from a center with a potential specialized interest in the outcome

Class IV

- Did not include persons at risk for the outcome
- Did not statistically sample patients or patients specifically selected for inclusion by outcome
- Undefined or unaccepted screening procedure or outcome measure
- No measure of frequency or statistical precision calculable

Appendix 5: Sample Evidence Tables

Design Characteristics and Outcomes in Controlled Studies of Patients with Bell's Palsy Treated with Steroids

Author Year	Class	Blind	Cohort Size	Completion Sate %	Steroid Dose Duration Rx	Follow-up Months	Severity %	Duration Days	NH %	RR Good Recovery (CI)	RR Complete Recovery (CI)
May 1976[7]	I	Yes	51	100	Prednisone 410 mg 10 days	6	47	2	81	0.99 (0.76-1.30)	0.92 (0.60-1.4)
Taverner 1954[8]	I	Yes	26	100	Hydrocortisone 1 gm 8 days	NS	23	9	67	1.07 (0.64-1.80)	–
Brown 1982[9]	I	Yes	82	100	Unnamed 400 mg 10 days	12	0	3	73	1.20 (0.97-1.50)	1.20 (0.97-1.49)
Wolf 1978[10]	I	No	239	100	Prednisone 760 mg 17 days	12	31	5	98	1.02 (0.99-1.06)	1.09 (0.98-1.22)
Austin 1993[11]	I	Yes	76	71	Prednisone 405 mg 10 days	6	22	5	83	1.21 (1.05-1.39)	1.71 (1.00-2.95)
Shafshak 1994[12]	II	Yes	160	100	Prednisolone 420 mg 10 days	12	91	6	69	1.24 (1.03-1.49)	1.76 (1.08-2.87)
Adour 1972[6]	II	No	304	85	Prednisone 216 mg 12 days	1	NS	14	64	1.39 1.20-1.62	1.58 (1.25-2.00)
Prescott 1988[13]	II	No	879	66	Prednisolone 520 mg 8 days	9	51	7+	92	1.04 (0.99-1.09)	1.04 (0.99-1.09)

Completion rate: Percentage of subjects followed to study completion. Severity: Percentage of patients with complete palsy. Duration: Maximum duration of palsy before starting steroids.

NH: Natural history, percentage of non–steroid-treated patients attaining a good outcome. RR: relative rate of steroid-treated patients attaining outcome compared to non–steroid-treated patients.

CI: 95% confidence intervals. NS: Not stated.

Design Characteristics and Outcomes in Controlled Studies of Patients with Bell's Palsy Treated with Acyclovir

Author Year	Class	Blind	Cohort Size	Completion Rate %	Dose Duration Rx	Follow-up Months	Severity %	Duration Days	NH %	RR Good Recovery (CI)	RR Complete Recovery (CI)
Adour 1996[15]	I	Yes	99	83	400 mg x 5 qd 10 days	12	20	3	76	1.22 (1.02-1.45)	1.21 (0.98-1.49)
De Diego 1998[16]	I	No	101	89	800 mg tid 10 days	3	1	4	94	0.83 (0.71-0.98)	–
Ramos 1992[17]	I	No	30	100	1000 mg qd 5 days	NS	63	NS	100	1.00*	–

Completion rate: Percentage of subjects followed to study completion. Severity: Percentage of patients with complete palsy. Duration: Maximum duration of palsy before starting steroids. NH: Natural history, percentage of non–acyclovir-treated patients attaining a good outcome. RR: relative rate of acyclovir-treated patients attaining outcome compared to non–acyclovir-treated patients. CI: 95% confidence intervals.

NS: Not stated. *All patients with good recovery.

Design Characteristics and Outcomes in Controlled Studies of Patients with Bell's Palsy Treated with Facial Nerve Decompression

Author Year	Class	Blind	Cohort Size	Completion Rate %	Surgical Approach	Follow-up Months	Severity %	Duration Days	Nh %	RR Good Recovery (CI)	RR Complete Recovery (CI)
Brown 1982[9]	II	No	92	100	Vertical, stylomastoid, midcranial fossa	12	100	14	47	1.21 (0.97-1.5)	1.30 (0.89-1.90)
Gantz 1999[18]	II	No	70	100	Midcranial fossa & meatal foramen	7	100	14	42	2.19	2.96
May 1981[19]	II	No	60	100	Transmastoid, vertical	6	92	14	6	1.14 (0.79-1.65)	6.4 (0.92-45)
May 1985[20]	II	No	38	100	Transmastoid, extralabyrinthine, subtemporal	6	100	14	23	0.87 (0.24-3.07)	–
Fisch 1981[21]	II	No	27	100	Midcranial fossa & meatal foramen	12-36	100	21	15	3.30 (0.82-12.90)	–

Completion rate: Percentage of subjects followed to study completion. Severity: Percentage of patients with complete palsy. Duration: Maximum duration of palsy before starting steroids. NH: Natural history, percentage of nonsurgical patients attaining a good outcome. RR: relative rate of surgically treated patients attaining outcome to non–surgically treated patients. CI: 95% confidence intervals. NS: Not stated.

Appendix 6: Tools for Building Conclusions and Recommendations

The following tools are provided to assist in the development of conclusions and recommendations. Keep in mind that all mandatory elements must be included (or obviously implied) in each conclusion and recommendation statement. The exact wording and order of the elements can vary from those suggested for grammatical and stylistic considerations. For the wording of conclusions when there is insufficient evidence see the examples below. The examples are hypothetical.

Mandatory Elements	Suggested Verbiage
Patient Population:	**For Patients with Condition X it is**
Strength of evidence (pick one): - strong - moderately strong - weak - insufficient	 - highly likely (highly probable) that - likely (probable) that - possible that - insufficient evidence to support or refute that
Intervention (co-intervention):	Intervention A (as compared with intervention B)
Effect (pick one): - Therapy/Causation - Prognosis/Diagnosis/Screening	 - is (not) effective in reducing the risk of - (does not) increase(s) the risk of - is (not) useful (predictive) in identifying - patients at increased risk for - patients with - a (treatable important) cause of
Outcome:	Outcome Y (if possible include a magnitude of effect)
Evidence summary:	(Number of studies and their Class)

Examples

Therapy

For patients with Bell's palsy it is highly likely that prednisolone (as compared with placebo) is effective in reducing the risk of incomplete facial functional recovery—risk reduction 12% (two Class I studies).

For patients with Bell's palsy it is highly likely that antivirals (as compared with placebo) are not effective in reducing the risk of incomplete facial functional recovery (two Class I studies).

Causation

For persons at risk for developing multiple sclerosis (MS) it is possible that low serum vitamin D levels increase the risk of the development of MS—odds ratio 1.23 (two Class III studies).

For young children it is likely that immunizations do not increase the risk of autism (multiple Class II studies).

Diagnostic Accuracy

For patients with rapidly progressing dementia it is likely that CSF 14-3-3 assays are useful in identifying patients with prion disease—sensitivity 80%, specificity 85% (multiple Class II studies).

For patients with symptoms and signs suggesting carpal tunnel syndrome it is highly likely that the flick sign is not useful in identifying patients with carpal tunnel syndrome—sensitivity 80%, specificity 20% (multiple Class I studies).

Prognostic Accuracy

For patients with cryptogenic ischemic stroke it is possible that the presence of patent foramen ovale (PFO) is useful in identifying patients at increased risk of recurrent ischemic stroke (two Class III studies).

For patients with ischemic stroke it is highly likely that elevated serum homocysteine levels identify patients at increased risk of recurrent stroke—relative risk 1.6 (two Class I studies).

Population Screening

For children with global developmental delay (GDD) it is possible that routine MRI of the head is useful in identifying a cause of the GDD—yield 4.5% (multiple Class III studies).

For patients meeting International Headache Society (IHS) criteria for migraine and a normal neurologic examination it is likely that routine head imaging (MRI or CT) is not useful in identifying important abnormalities—yield 0.5% (single Class I study).

Insufficient Evidence

For patients with Alzheimer's disease there is insufficient evidence to support or refute the effectiveness of coenzyme Q for slowing cognitive decline (Class IV studies only).

For patients with post–cardiac arrest brain anoxia there is insufficient evidence to support or refute the usefulness of visual evoked potentials in identifying patients at low risk of recovery (inadequately powered Class II study).

Recommendations

Elements of Recommendations

Mandatory Elements	Suggested Verbiage		
When (in what circumstances and in what patient population)	(For/In) patients with condition X		
Who (the person performing the action of the recommendation statement)	Clinicians		
Level of obligation (A, B, C)	A: Must (not) prescribe, offer (Rx) Must (not) test, counsel, monitor (Scrn, Dx, Px) Must avoid (causation)	B: Should (not) offer, prescribe Should (not) test, counsel, monitor Should avoid	C: May offer, prescribe May test, counsel, monitor, educate* May avoid May choose not to offer, prescribe May choose not to test, counsel, monitor
What (do what): Intervention (co-intervention): Intervention A (as compared with intervention B)	Describe specific intervention/test		
To precipitate what: (outcome)	Outcome Y		
Level of evidence: (Level N)			

*In the special case of negative Level C recommendations, we add the word *choose* because the term *may not* connotes a higher level of obligation than is intended.

Examples

Many of these examples would need specific clinical context sections to explain the rationale behind the recommendation and also any issues of generalizability, cost, etc.

Therapy

In patients with new-onset Bell's palsy, clinicians must prescribe prednisolone to reduce the risk of incomplete facial functional recovery (Level A).

In patients with Bell's palsy, clinicians must not offer antivirals (as compared with placebo) to reduce the risk of incomplete facial functional recovery (Level A).

Clinicians should offer patients with mild to moderate Alzheimer's disease cholinesterase inhibitors to slow the rate of cognitive decline (Level B).

For patients with mild to moderate Alzheimer's disease, clinicians should offer cholinesterase inhibitors to modestly slow the rate of cognitive decline (Level B).

Clinicians may choose not to offer mycophenolate to patients with generalized myasthenia gravis who are taking steroids to allow more rapid tapering of steroids (Level C).

Causation

Persons at risk for developing MS may avoid low serum vitamin D levels to decrease their risk of developing MS—odds ratio 1.23 (Level C).

Parents and clinicians should not avoid immunizations in young children to decrease the risk of autism (Level B).

Diagnostic Accuracy

Clinicians must inform families and patients with rapidly progressing dementia that the presence of CSF 14-3-3 protein increases the likelihood of prion disease (Level A).

Clinicians should inform families and patients with rapidly progressing dementia that the presence of CSF 14-3-3 protein increases the likelihood of prion disease (Level B).

Clinicians may choose to inform families and patients with rapidly progressing dementia that the presence of CSF 14-3-3 protein increases the likelihood of prion disease (Level C).

Clinicians may choose not to inform patients with symptoms and signs suggesting carpal tunnel syndrome with a flick sign that they are more likely to have carpal tunnel syndrome (Level C).

Prognostic Accuracy

Patients with cryptogenic ischemic stroke should be counseled that the presence of PFO is not useful in identifying patients at increased risk of recurrent ischemic stroke (Level B).

Patients with cryptogenic ischemic stroke may be counseled that the presence of PFO is not useful in identifying patients at increased risk of recurrent ischemic stroke (Level C).

Clinicians should inform patients with cryptogenic stroke that the presence of a PFO does not increase their risk of subsequent stroke (Level B).

Clinicians must inform patients with cryptogenic stroke that the presence of a PFO does not increase their risk of subsequent stroke (Level A).

Clinicians must not inform patients with cryptogenic stroke that the presence of a PFO increases their risk of subsequent stroke (Level A).

Population Screening

For children with GDD, clinicians may order routine MRI of the head to identify a cause of the GDD (Level C).

Clinicians may offer MRI of the head to children with GDD to identify the cause of the GDD (Level C).

Clinicians should not offer head imaging to patients meeting IHS criteria for migraine and a normal neurologic examination to identify important abnormalities (Level B).

Clinicians should not routinely perform head imaging for patients meeting IHS criteria for migraine and a normal neurologic examination to identify important abnormalities (Level B).

Insufficient Evidence

For patients with Alzheimer's disease, there is insufficient evidence to make recommendations regarding the use of coenzyme Q for slowing cognitive decline (Level U).

For patients with post–cardiac arrest brain anoxia, there is insufficient evidence to make recommendations regarding the usefulness of visual evoked potentials in identifying patients at low risk of recovery (Level U).

Appendix 7: Clinical Contextual Profile Tool

Recommendation Level	Level U	Level C	Level B	Level A
Wording	None	May	Should	Must
Adherence expected to affect	Few	Some	Most	Nearly all
Variation in patient preferences	Large			Minimal
Cost	Prohibitive			Minimal
Availability	Limited			Universal
Value of benefit relative to risk	Too close to call	Small	Moderate	Large
Confidence in evidence	Very Low	Low	Moderate	High
Strength of principle-based inferences	Not plausible	Plausible	Convincing	Compelling

Recommendation level anchored by the lowest of the confidence-in-evidence and the strength of principle-based inferences. The recommendation level can be decreased for any factor. The recommendation level can be increased only for a value-of-the-benefit relative to risk that is moderate or large and can only increase by one level. With the exception of the unusual circumstance of recommendations derived solely form first principles, the level of recommendation can never attain Level A without "High" evidence.

Appendix 8: Conflict of Interest Statement

Definitions of Terms in Disclosure Agreement

Commercial entity: A for-profit business that manufactures, distributes, markets, sells, or advertises pharmaceutical or scientific products or medical devices.

Compensation: Anything of monetary value including a salary, honorarium, stipend, gift, or payment of travel-related expenses.

Expert Witness: A person who has provided expert medical testimony during a trial or administrative hearing, in a deposition or an affidavit, or in any other type of legal proceeding.

Immediate "Family Member": Any person who would benefit financially from the publication of the manuscript because of their relationship to the author. This includes a member of an author's immediate family or anyone else who has a significant relationship with the author.

I have read and understand the definitions of the above terms.

Name: _____

Guideline project(s):_____

Nonfinancial Disclosure	
1. I take responsibility for the contributions previously indicated and the conduct of the research. I had full access to the data.	
Yes	
No	
2. I have chosen to declare one or more **non-financial** competing interests (e.g., special interest groups you represent or others that may be affected if your paper is published or **that could be perceived as biasing the study**). Non-financial disclosures will not be published.	
Yes	
No	

Financial Disclosure	
Personal Compensation from a Commercial or Non-Profit Entity **Within the past 24 months (and during the course of the study under consideration if the study exceeded two years), I or one of my "immediate family members" received personal compensation for the following:** (All compensation received during the past two years regardless of the relationship to the study must be disclosed; for the period exceeding two years, only compensation relevant to the topic of the study needs to be disclosed.)	
3. Serving on a scientific advisory board	
Yes	
No	
4. Gifts worth more than $500 (specify gift item and source in box)	
Yes	
No	
5. Funding for travel	
Yes	
No	
6. Serving as a journal editor, an associate editor, or as a member of an editorial advisory board. This may include a journal published by your national medical/scientific organization.	
Yes	
No	
7. Patents held or pending that may accrue revenue, whether or not revenue has been received to date	
Yes	
No	

8. Royalties from publishing	
Yes	
No	
9. Honoraria	
Yes	
No	
10. Corporate appointments	
Yes	
No	
11. Speakers' bureau	
Yes	
No	
12. Other non-CME related activities not covered in designations above	
Yes	
No	
13. Do you perform clinical procedures or imaging studies in your practice that overlap with the content of **this study, practice parameter, or clinical practice guideline** and would this part of your practice grow if the conclusions were widely followed? (**Note**: This is the only item in this Agreement that applies to an interest that is related **specifically** to this particular study, practice parameter, or clinical practice guideline.) If yes, provide details. Types of procedures and percentage of clinical effort (e.g. MRI – 25%)	
Yes	
No	

Research Support
Within the past 24 months and during the course of the study under consideration if the study exceeded two years, I or one of my "immediate family members" received financial or material research support or compensation from the following.
(All support received during the past two years regardless of the relationship to the study must be disclosed; for the period exceeding two years, only support relevant to the topic of the study needs to be disclosed.)
14. Commercial entities
Yes
No
15. Government entities
Yes
No
16. Academic entities other than those attributed in the manuscript
Yes
No

Stock, Stock Options, and Royalties
In the past 24 months and during the course of the study under consideration if the study exceeded two years, I or one of my "immediate family members": (All revenues during the past two years regardless of the relationship to the study must be disclosed; for the period exceeding two years, only revenues relevant to the topic of the study needs to be disclosed.)
17. Held stock or stock options or received expense compensation for serving on a board of directors
Yes
No

18. Received license fee payments	
Yes	
No	
19. Received royalty payments or have contractual rights for receipt of future royalty payments from technology or inventions that have been licensed or sold (this does not include royalties from publishing).	
Yes	
No	
20. Held stock or stock options in a company sponsoring research with which the author or "immediate family member" was involved as an investigator. (This excludes investments in mutual funds held by the author or dependents.)	
Yes	
No	
21. Held stock options in a company whose medical equipment or other materials related to the practice of medicine. (This excludes investments in mutual funds held by the author or dependents.)	
Yes	
No	

Legal Proceedings	
In the past 24 months and during the course of the study under consideration if the study exceeded two years, I have (whether or not it pertains to the topic of the current study):	
22. Given expert testimony with regard to any legal proceeding	
Yes	
No	
23. Prepared an affidavit with regard to any legal proceeding	
Yes	
No	
24. Acted as a witness or consultant with regard to any legal proceeding	
Yes	
No	

I have completed this Disclosure Statement fully and to the best of my ability. I understand all Authors must complete this Disclosure Statement and that the information disclosed will be published if the manuscript is accepted for publication.

Signed: _____

Date: _____

Appendix 9: Project Development Plan Worksheet

1. Clinical Question Development:

a. Problem/Issue to be addressed:

b. To what patient population does this apply?

c. What is the intervention (therapy, test, risk factor)?

d. What are the outcomes of interest?

e. State one or more answerable clinical questions that include the population, intervention, and outcomes of interest:

Examples:

- *What is (are) the best medication(s) for controlling seizures while minimizing side effects and providing a good quality of life for a patient who requires treatment for epilepsy?*
- *Does anticonvulsant prophylaxis decrease the risk of developing late seizures in patients with head injury?*
- *In patients with Bell's palsy, do steroids improve facial function outcomes?*

2. Criteria for Literature Search:

a. Key Text words and Index words for the condition or closely related conditions, if appropriate (linked by the word "OR"): _____

b. Key Text words and Index words for the intervention (linked to above by the word "AND"): _____

c. Databases to be searched (e.g. MEDLINE, EMBASE, Current Contents): _____

d. Years to be included in the search: _____

3. Inclusion and Exclusion Criteria:

a. Include all languages: ☐ Yes ☐ No

b. Selected study population:

Human subjects: ☐ Yes ☐ No

Animal studies: ☐ Yes ☐ No

c. Disease in question or closely related diseases to be included: _____

d. Interventions to be included: _____ Interventions to be excluded: _____

e. Outcomes to be included: _____ Outcomes to be excluded: _____

f. Types of studies to be included:

☐ RCT ☐ Cohort ☐ Case control ☐ Case series (☐ must be greater than _____

☐ Review papers ☐ Meta-analyses

g. Standard exclusion criteria:

- Not relevant to the clinical question
- Unrelated disease
- Outside of study population
- Article not peer reviewed

h. Additional exclusion criteria:

_____

_____

4. Project Timeline (enter dates based on your availability and the guidelines provided)

- Complete panel formation by _____ (usually takes two to four weeks)
- Literature search _____ (select a timeframe of one to two weeks, during which you will have time to complete the search with the librarian and review and distribute the abstracts; AAN staff will have the librarian contact you to begin this step)
- Panel review of literature _____ (two-step process of reviewing abstracts and then selected articles – takes six to eight weeks)
- Data extraction and development of evidence tables _____ (takes three to eight weeks depending on total number of articles to analyze and tabulate)
- Drafting the guideline _____ (takes four to eight weeks)
- Goal for submitting first draft to GDS _____

GDS: Months in which drafts usually accepted (select one): ☐ March ☐ June ☐ September ☐ December

Year _____

Appendix 10: Sample Data Extraction Forms

Atrial Fibrillation Rx Data Extraction Form-DRAFT

Patient Population: For patient with nonvalvular atrial fibrillation (Including these special populations: Patients with intracranial hemorrhages (spontaneous transformation, posttraumatic, hypertensive, vascular malformation); Patients with intracranial or intraspinal (vascular malformations); Patients s/p CABG; other special populations (Elderly, nursing home residents, end-stage renal disease, dementia)

Intervention: What therapies (Including: Antithrombotics: Warfarin, aspirin, dabigatran, apixaban, rivaroxaban, combination therapy; Rate or rhythm control of atrial fibrillation: with medical therapy or ablation)

Comparative Intervention: Compared with no therapy or another therapy

Outcomes: Reduce the risk of ischemic stroke with the least risk of hemorrhage (including intracerebral hemorrhage)

Summary

Panel member: _____

Article ID#: _____

Inclusion Criteria

- Human studies only
- Enrolled patients with atrial fibrillation
- Patients receiving different therapies to prevent ischemic stroke

Exclusion Criteria

- Case report, editorial, meta-analysis, or review (please specify)
- < 50 patients

Comparison Group

- The study compares outcomes between groups using different management strategies (e.g., ablation plus anticoagulation to anticoagulation alone).
- To be considered a randomized controlled trial, patients should have been randomized to different management strategies.

Relevance

Study is relevant to question? ☐ Yes ☐ No

If no, STOP. Explain _____

Design

☐ Randomized controlled trial

☐ Nonrandomized trial that includes a comparison group

If the study does not include a comparison group, STOP.

(The study does not meet inclusion criteria)

For the therapeutic rating:

If a randomized controlled trial, maximum Class I.

If not a randomized trial, maximum Class II.

▶ **MAXIMUM THERAPEUTIC CLASS** ☐ I ☐ II

Sample Size

Total patients enrolled: _____

If total less than 50 patients, STOP (Study does not meet inclusion criteria)

Outcome Assessment

1. Was any outcome assessment blinded to management strategy?
☐ Yes ☐ No ☐ Not stated

2. Was any outcome objective?
☐ Yes ☐ No ☐ Not stated

3. Was any outcome assessed independently?
☐ Yes ☐ No ☐ Not stated

Comments regarding outcome assessment:
If 1 or 2 = YES, maximum is Class I
If only 3 = YES, maximum is Class III
If all = NO/NOT STATED, STOP: Class IV

▶ **MAXIMUM THERAPEUTIC CLASS** ☐ I ☐ II ☐ III ☐ IV

Outcomes

- Ischemic Stroke
- Bleeding

Objective

The determination of the outcome is unlikely to be affected by observer expectations. Consider the following outcomes objective: Death, Disabling Stroke, Major hemorrhage.

Independently

The investigator determining the outcome was different than the treating physicians.

Other Therapeutic Study Characteristics

1. Was treatment allocation concealed (Check "no" if not an RCT)
☐ Yes ☐ No ☐ Not stated

2. Primary outcome measure(s) was specified
☐ Yes ☐ No ☐ Not stated
Record primary outcome(s) _____
Record secondary outcomes _____

3. Explicit inclusion and exclusion criteria were used
☐ Yes ☐ No ☐ Not stated
Summarize relevant criteria _____

4. Patients in different treatment arms were similar at baseline or appropriate statistical adjustments were made for baseline differences
☐ Yes ☐ No ☐ Not stated

5. Less than 20% of patients were lost to follow-up
☐ Yes ☐ No ☐ Not stated

Percentage lost to follow-up:
If all = "yes," maximum is Class I.
If only three or four = "yes," maximum is Class II.
If < three = "yes," maximum is Class III.

▶ **MAXIMUM THERAPEUTIC CLASS** ☐ I ☐ II ☐ III ☐ IV

"Concealed Allocation"

Investigators could not manipulate treatment assignment. Examples of concealed allocation include consecutively numbered sealed, opaque envelopes containing a predetermined, random sequence for treatment assignment or an independent center that an investigator contacts to obtain the treatment assignment.

Final Rating: Select worst maximum therapeutic class from above

▶☐I ☐II ☐III ☐IV

If CLASS IV, STOP

Demographics (for entire study population if possible. Otherwise list values for all groups)

Age: Central tendency: ☐Mean ☐Median

Value: _____

Dispersion: ☐SD ☐SE ☐Range ☐Interquartile range

Value: _____

Gender % female: _____

Special Atrial Fibrillation Populations Included (check all that apply; describe)

☐Patients with intracranial hemorrhages (spontaneous transformation, posttraumatic, hypertensive, vascular malformation)
☐Elderly
☐Nursing home residents
☐End-stage renal disease
☐Dementia
☐Other _____
☐Other _____
☐Other _____
☐Other _____

Type(s) of Management Strategies (check all that apply; describe)

☐Aspirin
☐Clopidogrel
☐Clopidogrel plus aspirin
☐Warfarin
☐Dabigatran
☐Apixaban
☐Rivaroxaban
☐Triflusal &Warfarin
☐Medication(s) for rate or rhythm control
☐Ablation for rate or rhythm control
☐Other _____
☐Other _____
☐Other _____
☐Other _____

Describe Management Strategy Comparison Groups, Including the Number in Each Group (there should be at least two)

	Number	**Description of Group**
Group 1	_____	_____
Group 2	_____	_____
Group 3	_____	_____
Group 4	_____	_____

Outcomes Described

Thromboembolic Events (check all described)

- ☐ Ischemic stroke
- ☐ TIA
- ☐ All ischemic stroke
- ☐ Fatal ischemic stroke
- ☐ Disabling ischemic stroke
- ☐ Nondisabling ischemic stroke
- ☐ Other _____
- ☐ Other _____
- ☐ Other _____

Comments: _____

Bleeding Events (check all described)

- ☐ Minor bleeding
- ☐ Major bleeding
- ☐ Intracranial bleeding
- ☐ Death secondary to hemorrhage
- ☐ GI hemorrhage
- ☐ Other bleeding events
- ☐ Other bleeding events

Other Outcomes (check all described)

- ☐ All-cause death
- ☐ Other _____
- ☐ Other _____
- ☐ Other _____
- ☐ Other _____

Results (Briefly summarize the study's results)

Comments (Provide any special reasons to include, noteworthy findings, reason for classifying, etc.).

Atrial Fibrillation Population Screening DRAFT

For patients with cryptogenic stroke how often do various technologies (including inpatient telemetry, Holter monitor, implanted recorder EKG) (as compared with not using the technology) identify patients with unsuspected atrial fibrillation (a. fib.) who would benefit from prophylaxis?

Inclusion Criteria
Study enrolls a series of ischemic stroke patients without known atrial fibrillation.
The population is evaluated by some technique (e.g., inpatient telemetry) for atrial fibrillation.
The study enumerates the proportion of patients identified with atrial fibrillation.

Panel member: _____

Article ID#: _____

Relevant to question? ☐ Yes ☐ No

If NO, STOP. Explain _____

Part A: Class of Evidence

Were stroke patients with a. fib. specifically recruited for the study?

☐ Yes ☐ No ☐ Not stated/uncertain

If yes, STOP, study is Class IV.

Sample Size

Total subjects enrolled: _____

If total less than 50 subjects, STOP (Study does not meet inclusion criteria)

Major Study Characteristics

From where were the subjects recruited for the study (Sampling Frame)? Select one.

- ☐ A defined geographic area (i.e., population-based, sampling frame includes all ischemic stroke patients without known a. fib. within the geographic region). *Maximum Class I*
- ☐ A general medical, neurology clinic/hospital without a specialized interest in a. fib. and stroke. *Maximum Class II*
- ☐ A referral center/clinic with a specialized interest in a. fib. and stroke. *Maximum Class III*

▶ **MAXIMUM CLASS** ☐ I ☐ II ☐ III

How were stroke patients selected for inclusion in the study?

- ☐ All subjects in the sampling frame were invited/recruited into the study (consecutive). *Maximum Class I*
- ☐ A random or systematically selected (e.g., every 3rd patient) subset of subjects were invited/recruited into the study. *Maximum Class I*
- ☐ Nonsystematically selected subjects were included in the study. *Maximum Class III*
- ☐ Not stated/Uncertain. *Maximum Class III*

▶ **MAXIMUM CLASS** ☐ I ☐ II ☐ III

How many of the patients selected for inclusion in the study were actually screened for a. fib.?

- ☐ ≥80% screened for a.fib. *Maximum Class I*
- ☐ 50 to <80% screened for a.fib. *Maximum Class II*
- ☐ <50% screened for a.fib. *Maximum Class III*

▶ **MAXIMUM CLASS** ☐ I ☐ II ☐ III

Final Rating (select worst maximum Class)

▶ ☐ I ☐ II ☐ III

Part B: Study Details

Participant Characteristics

Describe the subjects included in the study.

Demographics (for entire study population if possible; otherwise describe for all subgroups)

Age: Central tendency: ☐ Mean ☐ Median

Value: _____

Dispersion: ☐ SD ☐ SE ☐ Range ☐ Interquartile range

Value: _____

Gender % female: _____

Presence of Stroke

What criteria were used to diagnose stroke: _____

Describe the ischemic stroke patients included: _____

Diagnostic Technique employed to look for a. fib. (describe, include duration).

- ☐ Inpatient telemetry _____
- ☐ Holter monitor _____
- ☐ Implanted recorder _____
- ☐ Other _____

Results. Briefly describe the results of the study:

Total number of patients studied for a. fib.: _____

Number of patients found to have a. fib.: _____

Comments (special reasons to include, noteworthy findings, etc.)

_____

_____

_____

Appendix 11: Manuscript Format

Cover Page

Evidence-based guideline (update): Title

Report of the Guideline Development Subcommittee of the American Academy of Neurology

List authors' names, designations, and institutional affiliations

Address correspondence and reprint requests to: American Academy of Neurology, 1080 Montreal Ave, St. Paul, MN 55116, *guidelines@aan.com*

Word counts for abstract and manuscript (includes only body of manuscript, excluding appendices and references)

Character count for title (including spaces)

Page 2

Conflict of interest disclosures

Manuscript

Abstract

Up to 250 words; should summarize the guideline as follows:

Objective: Summary of clinical focus

Methods: Description of process

Results/Conclusions: Status, quality, and content of evidence

Recommendations: Summary of recommendations

Introduction

The Introduction should concisely cover the following:

- Statement of Purpose (including identification of audiences)
- Background and Justification. An overview of the problem or topic area under study and the underlying justification for pursuing the question. May include any or all of the following:
  - Membership needs; the degree of interest and usefulness to Academy members, if known (e.g., by survey)
  - The potential for significant benefit or risk to patients and abuse
  - Extent of practice variation
  - Urgency
  - Controversy regarding validity or applicability
- Clinical Question Statement

Description of the Analytic Process

This section should present the exact, replicable process the authors used to develop the guideline, including:

- How the panel was selected, including disclosure of information, funding, and outside input (e.g., reviewers)
- Description of literature review
  - How the literature search was conducted (search terms, databases searched, other search strategies, languages included, dates covered). Describe bibliographic or other search techniques in sufficient detail so that the process can be replicated.
  - How articles were selected for inclusion (e.g., all articles reviewed, only prospective studies selected, etc.).
    - Inclusion and exclusion criteria and process for "weeding out" articles
    - State the number of articles identified in the search, the number excluded during the abstract review, the number excluded during the article review, and the number eventually included in the guideline.
    - State how abstracts and articles were reviewed (e.g., how many panel members reviewed each, how disagreements were resolved)
  - Analysis of the data
    - Elements of evidence extracted from pertinent articles, using a data extraction form
    - Classification of evidence definitions
    - Development of evidence tables

Analysis of Evidence

This section is the scientific body of the paper and should include a detailed narrative description of the evidence and the statistical analysis applied to it, as appropriate to the topic. If more than one clinical question is addressed, it is appropriate to deal with the questions one at a time, providing data analyzed, levels of evidence, conclusions, and recommendations for each question.

For diagnostic tests:

- Results
- Levels of evidence
- Statistical analysis (meta-analysis, sensitivity and specificity, positive and negative predictive values, ORs, relative rates, and numbers needed to treat/harm)
- Relevance (selection criteria, complications, contraindications, test specifics)
- Clinical significance
- Availability of a reference standard (gold standard) for comparison

For therapies:

- Results
- Levels of evidence
- Statistical analysis (meta-analysis, sensitivity and specificity, positive and negative predictive values, ORs, relative rates, and numbers needed to treat/harm)
- Relevance (patient selection criteria, complications, contraindications, intervention details, protocols, difficulty with implementation, duration/frequency of treatment)
- Clinical significance

Conclusions

This section summarizes the evidence in answer to the clinical question. The conclusions should be directly linked to the evidence (e.g., Four Class II studies show…).

Clinical Context

This is an optional section providing information regarding alternatives for which there was limited evidence, risk/benefit profiles, limits to the generalizability of the evidence, magnitude of benefit, harms, cost, outcomes not addressed in the evidence, etc. In short, this section may include any information that does not directly follow from the evidence presented. This section can be presented after the conclusions section of each clinical question (as needed).

Recommendations

This section translates the conclusions into action statements. Each recommendation must be clearly linked to the evidence and include a quality of evidence label (e.g., Level A). Recommendations should not be broader or narrower than the clinical question.

Recommendations for Future Research

This section presents the identified gaps in the literature.

Tables/Figures

Tables, algorithms, or figures should be presented if they help communicate—but not alter—the evidence-based recommendations. In most cases, evidence tables are placed online.

Disclaimer

The following disclaimer must appear on all guidelines:

> This statement is provided as an educational service of the American Academy of Neurology. It is based on an assessment of current scientific and clinical information. It is not intended to include all possible proper methods of care for a particular neurologic problem or all legitimate criteria for choosing to use a specific procedure. Neither is it intended to exclude any reasonable alternative methodologies. The AAN recognizes that specific patient care decisions are the prerogative of the patient and the physician caring for the patient, and are based on all of the circumstances involved.

Conflict of Interest Statement

Acknowledgments

Appendices

Appendices will include GDS members and the schemes for classification of evidence and classification of recommendations. This section can be populated by AAN staff, and parts of it may be placed online.

References

References up to and including 40 will be included in the print version of the journal, and those beyond 40 will be published online.

Appendix 12: Sample Revision Table

#	Reviewer	Criticism	Action
1	A.B. Smith	1. Clarify the diagnostic criteria	1. A sentence has been inserted about diagnostic criteria citing the World Federation of Neurology criteria
		2. PEJ vs PEG	2. There is little evidence on PEJ and expert consensus was not achieved – no action
		3. "Breaking the News" is a flippant term	3. No change; the term was derived from the literature and from consensus of the task force
		4. Editorial changes suggested	4. Selectively incorporated
2	X.Y. Jones	1. Many aspects of symptomatic care are not covered	1. No change; to be covered in future practice parameters
		2. Some evidence from only 1 or 2 studies provides the basis for some recommendations, e.g., sialorrhea	2. No change; this is the status of the evidence
		3. We omitted data from Belsch and Shipman in a book chapter	3. No change; reference not added since no measures of quality of life or survival were made
		4. The recommendation about invasive ventilation should be separated and expanded to include fully informing about burdens and benefits	4. So changed
3	Anonymous	1. Delete the option on laryngectomy for recurrent aspiration	1. No change; evidence supports its consideration in patients with both aphonia and recurrent aspiration
		2. The word *entrapment* with respect to tracheostomy/ ventilator without proper planning is unclear	2. The word *entrapment* is dropped and the phrase clarified
		3. Extensive editing	3. Selectively accepted