

Advanced/Policy Track

Evidence-Based Guidelines

By Paul H Barrett, MD, MSPH
Grant Okawa, MD
Jill Bowman, BS

Abstract

The Advanced/Policy Track of the 2004 Kaiser Permanente Evidence-Based Medicine Symposium was an interactive session that focused on developing evidence-based clinical practice guidelines. The hypothetical scenario involved the imaginary drug "Memoryboost," a treatment for dementia. The participants were given materials describing the national Kaiser Permanente (KP) methodology for developing evidence-based guidelines and a summary of the highest-quality articles about the efficacy of this drug. The participants then formed small groups and used this information to develop a recommendation about its use for the treatment of dementia. In spite of having the same evidence, the groups developed three different recommendations. The entire group then explored some of the reasons for this variability. This article also addresses the reasons KP develops its own national guidelines, as well as who oversees the national guideline initiative and who develops guidelines.

Introduction

The Advanced/Policy Track was an interactive session that focused on developing evidence-based clinical practice guidelines. The scenario involved the fictitious drug "Memoryboost," a treatment for dementia. In this article we describe the National Kaiser Permanente (KP) Guideline process and then review the hypothetical "Memoryboost" case study as an example of the application of that process.

KP National Guidelines

KP has developed interregional guidelines for several years. The primary reasons that KP develops its own guidelines are the need for consistency across all of our regions and the economies of scale that our large size permits. KP needs con-

sistency because it is expected from our large national accounts and because the programs, systems, and materials necessary for the implementation of guidelines can be produced with higher quality and more economically by collaborating across the program. In addition, the members of the team that develops the guideline constitute a group of advocates in each medical group and region. These individuals serve as two-way conduits of information, taking local concerns and priorities to the team and then sharing draft guidelines with other clinical leaders in the regions for their input prior to publication.

With the exception of guidelines developed by rigorous evidence-based methods, eg, some guidelines developed by the US Preventive

Services Task Force, KP rarely adopts existing guidelines published by other groups, such as specialty societies or disease advocacy organizations, for several reasons. First, these groups may lack the necessary broad representation by the appropriate specialties, disciplines, and key stakeholders. Second, the KP evidence-based methodology (Common Methodology)¹ developed by an interregional collaboration of guideline experts, is more rigorous than that of many of these organizations. Third, the KP team members are less likely to have significant conflicts of interest with industry compared to national experts who sit on other guideline teams. Finally, KP, as an organization, has its own systems of care, formularies, and cost structure—all important factors to be taken into consideration when developing national guideline recommendations.

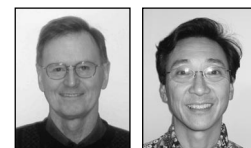
National KP guidelines are developed under the aegis of the National Guideline Directors, which includes the lead physicians in guideline development in each region. KP Guideline Development Teams (GDT) consist of members from all eight regions representing all the relevant health professions and specialties. Each team also includes an analyst and a physician trained in evidence-based methodology, who

... KP rarely adopts existing guidelines published by other groups ...

Paul H Barrett, MD, MSPH, (left) is an internist and Codirector of the KPCO Clinical Research Unit. He also chairs the KP Colorado Guideline Committee, the National KP Guideline Quality Committee and the National Clinical Content Network Review Board. E-mail: paul.h.barrett@kp.org.

Grant Okawa, MD, (right) is the Physician Lead for Ambulatory KP HealthConnect for the Hawaii Region. E-mail: grant.okawa@kp.org.

Jill Bowman, BS, (not pictured) was with the CMI's Evidence-Based Methodologist at the time of the KP EBM Symposium. Her role was to ensure that all of the CMI National Clinical Practice Guidelines met KP's rigorous EBM standards. Ms Bowman now works as the manager of Knowledge and Information for KP's Archimedes Group. E-mail: jill.d.bowman@kp.org.



Special Feature

Table 1. System for grading the strength of a body of evidence¹

Level/Grade	Therapy/Prevention/Screening	Diagnosis	Prognosis
Grade GOOD	<p>Type and number of studies</p> <ul style="list-style-type: none"> At least one well-designed, well-conducted systematic review (SR)/meta-analysis (MA) (consider heterogeneity) of RCTs Two or more well-designed, well-conducted RCTs with narrow confidence intervals One well-designed, well-conducted multi-center RCT with narrow confidence intervals <p>Quality</p> <ul style="list-style-type: none"> Low risk of bias Adequate sample size and power No major methodological concerns <p>Consistency</p> <ul style="list-style-type: none"> For SR/MA, no major conflict in results (consider heterogeneity). If significant heterogeneity exists, drops to Poor For individual RCTs, no major conflict in results If major conflicts do exist, drop to "Insufficient" <p>Relevancy</p> <ul style="list-style-type: none"> No compelling reason not to generalize the published work to the target KP population 	<p>Type and number of studies</p> <ul style="list-style-type: none"> At least one well-designed, well-conducted SR/MA (consider heterogeneity) of cross-sectional studies using independent gold standard Two or more well-designed, well-conducted cross-sectional studies using an independent gold standard <p>Quality</p> <ul style="list-style-type: none"> Low risk of (verification) bias Independent gold standard No major methodological concerns <p>Consistency</p> <ul style="list-style-type: none"> For SR/MA no major conflict in results (consider heterogeneity) For individual studies, consistent diagnostic accuracy <p>Relevancy</p> <ul style="list-style-type: none"> No compelling reason not to generalize the published work to the target KP population 	<p>Type and number of studies</p> <ul style="list-style-type: none"> At least one well-designed, well-conducted SR/MA (consider heterogeneity) of prospective cohort studies Two or more well-designed, well-conducted prospective cohort studies <p>Quality</p> <ul style="list-style-type: none"> Low risk of bias No major methodological concerns <p>Consistency</p> <ul style="list-style-type: none"> For SR/MA no major conflict in results (consider heterogeneity) For individual studies, consistent prognosis in similar populations <p>Relevancy</p> <ul style="list-style-type: none"> No compelling reason not to generalize the published work to the target KP population
Grade FAIR	<p>Type and number of studies</p> <ul style="list-style-type: none"> Single well-designed, well-conducted RCT with narrow confidence intervals Two or more RCTs of lower quality Well-designed, well-conducted SR/MA of cohort studies (consider heterogeneity) <p>Quality</p> <ul style="list-style-type: none"> Minor methodological concerns <p>Consistency</p> <ul style="list-style-type: none"> For SR/MA, no major conflict in results (consider heterogeneity) For individual studies, no major conflict in results If major conflicts do exist, drop to "Insufficient" <p>Relevancy</p> <ul style="list-style-type: none"> No compelling reason not to generalize the published work to the target KP population 	<p>Type and number of studies</p> <ul style="list-style-type: none"> Single well-designed, well-conducted cross-sectional study Two or more cross-sectional studies of lower quality Well-designed, well-conducted SR/MA of lower quality studies <p>Quality</p> <ul style="list-style-type: none"> Minor methodological concerns Independent gold standard <p>Consistency</p> <ul style="list-style-type: none"> For SR/MA, no major conflict in results (consider heterogeneity) For individual studies, no major conflict in results <p>Relevancy</p> <ul style="list-style-type: none"> No compelling reason not to generalize the published work to the target KP population 	<p>Type and number of studies</p> <ul style="list-style-type: none"> Single well-designed, well-conducted prospective cohort study Two or more prospective cohort studies of lower quality Well-designed, well-conducted SR/MA (consider heterogeneity) of either retrospective cohort studies or untreated control arms in RCTs <p>Quality</p> <ul style="list-style-type: none"> Minor methodological concerns <p>Consistency</p> <ul style="list-style-type: none"> For SR/MA, no major conflict in results (consider heterogeneity) For individual studies, no major conflict in results <p>Relevancy</p> <ul style="list-style-type: none"> No compelling reason not to generalize the published work to the target KP population
Grade INSUFFICIENT NOTE: Any evidence that fails to meet criteria for GOOD or FAIR evidence is considered to be INSUFFICIENT. Examples of insufficient evidence are provided for the different criteria.	<p>Type and number of studies</p> <ul style="list-style-type: none"> Single RCT of lower quality or insufficient size Cohort study <p>Quality</p> <ul style="list-style-type: none"> Major methodological concerns (ie, lack of concealed allocation, inadequate blinding, no ITT analysis) <p>Consistency</p> <ul style="list-style-type: none"> Studies that are well-designed, well-conducted (Good or Fair) but with major conflict in results SR/MA with major conflict in results (consider heterogeneity) <p>Relevancy</p> <ul style="list-style-type: none"> Compelling reasons why the results do not apply to the target KP population 	<p>Type and number of studies</p> <ul style="list-style-type: none"> Single cross-sectional study of lower quality Case-control study <p>Quality</p> <ul style="list-style-type: none"> Major methodological concerns (nonconsecutive, poor or non-independent gold standard) <p>Consistency</p> <ul style="list-style-type: none"> Studies that are well designed, well-conducted (Good or Fair) but with major conflict in results <p>Relevancy</p> <ul style="list-style-type: none"> Compelling reasons why the results do not apply to the target KP population 	<p>Type and number of studies</p> <ul style="list-style-type: none"> Single prospective cohort study of lower quality Retrospective cohort study Untreated control arm of RCT Case series <p>Quality</p> <ul style="list-style-type: none"> Major design or methodological concerns (sampling bias, high dropout, nonblinded outcome assessment, lack of adjustment for confounders) <p>Consistency</p> <ul style="list-style-type: none"> Studies that are well-designed, well-conducted (Good or Fair) but with major conflict in results <p>Relevancy</p> <ul style="list-style-type: none"> Compelling reasons why the results do not apply to the target KP population

Table 2. Evidence table: "Memoryboost" for treatment of Alzheimer's disease (AD)

Study, total n	Treatment groups size and drug	Study population	Results	Comments
Winblad et al, 2001 (RCT, double-blind) Follow-up: 52 weeks Initial n: 286 Final n: 192	Rx1 placebo (n = 144) Rx2 "Memoryboost" 5 mg/day for 28 days, then increased to 10 mg/day (n = 142)	<ul style="list-style-type: none"> • Possible or probable • Alzheimer's disease MMSE ^a score ≥ 10 and ≤ 26	MMSE ^a (Mean change): Rx1: -2.2 Rx2: -0.5 p < 0.001	Study funded by the pharmaceutical company that produces "Memoryboost"

^aMMSE: Mini mental status examination

are expert in searching, summarizing, and critically appraising the medical literature.

National KP guidelines are developed according to a rigorous, evidence-based methodology, which includes five steps: Problem Formulation, Evidence Search, Evidence Summary, Rationale, and Recommendation. Each guideline addresses several discrete topics (Problem Formulations), each with one or more Recommendations for care. The Evidence Search, Evidence Summary (often presented as Evidence Tables) and Rationale document the process of developing these Recommendations.

The KP methodology requires that the Evidence Search be comprehensive and fully documented. Depending on the quality and quantity of publications pertinent to a question a Recommendation may be evidence-based or consensus-based. Consensus-based Recommendations are developed in situations where the evidence is insufficient to support an evidence-based Recommendation, but a clinical question needs to be addressed. One example is the interval for certain cancer screening tests, where there is clear evidence about the effectiveness of screening but no studies directly comparing the effectiveness of different intervals. In this example, the recommendation to screen is evidence-based, but the recommended interval is consensus-based.

Each Problem Formulation is

based on a single clinical question, which includes four key components: **P**atient (population), **I**ntervention, **C**omparison and **O**utcome—PICO. For example, in the treatment of dementia, one important question is: "*How should pharmacological agents be used to treat cognitive and functional decline associated with dementia?*" In the hypothetical scenario, the four key components are:

P (Patient): Men and women with diagnosed dementia

I (Intervention): "Memoryboost"

C (Comparison): Placebo or other drugs used to treat dementia (including rivastigmine, hormone replacement therapy, statins, and others)

O (Outcome): Cognitive ability, functional ability, and others.

The Problem Formulation informs the Evidence Search, which specifies which databases were searched and what terms were used. All publications that address a given problem formulation are identified and then, based on commonly accepted standard criteria, all the articles that are relevant and of sufficient quality are summarized. For approaches to rating the quality of journal articles, see Table 1 and *The User's Guide to the Medical Literature*.² For questions about treatment, the publications are generally restricted to high quality randomized clinical trials (RCTs). For more details about the kinds of studies used to address questions other than treatment

choices see Table 1.

The articles selected in the evidence search are summarized in either text or table format. The summary typically includes the reference to the article, the selection criteria and number of subjects in each arm of the study, the treatments being compared, the results and relevant comments about biases and other threats to the validity of the study. (See Table 2, taken from the hypothetical case study).

The team makes two judgments as it reviews the evidence summary:

1) What is the quality and quantity of the evidence? and 2) What are the benefits and harms of the treatments included in the summary? Evidence that is of adequate quantity and quality is said to be "sufficient," and is further classified as either "good" or "fair."

Unlike guidelines from many other organizations, KP national guidelines always have a rationale statement that explicitly ties each recommendation to its supporting literature. For National KP guidelines the rationale typically includes a narrative summary of the evidence tables and the conclusion that follows. In simplified form a rationale might state, "Based on one systematic review and two more recent randomized clinical trials the guideline development team concludes that drug A is highly effective for the treatment of cognitive decline in dementia." Where there

... KP national guidelines always have a rationale statement that explicitly ties each recommendation to its supporting literature.

Table 3. Language of recommendations¹

Evidence-Based Recommendations
<p>Recommendation: A</p> <p>Language:^a The intervention is strongly recommended for eligible patients.</p> <p>Evidence: The intervention improves important health outcomes, based on good evidence, and the Guideline Development Team (GDT) concludes that benefits substantially outweigh harms and costs.</p>
<p>Recommendation: B</p> <p>Language:^a The intervention is recommended for eligible patients.</p> <p>Evidence: The intervention improves important health outcomes, based on 1) good evidence that benefits outweigh harms and costs; or 2) fair evidence that benefits substantially outweigh harms and costs.</p>
<p>Recommendation: C</p> <p>Language:^a No recommendation for or against routine provision of the intervention. (At the discretion of the GDT, the recommendation may use the language "option," but must list all the equivalent options.)</p> <p>Evidence: Evidence is sufficient to determine the benefits, harms, and costs of an intervention, and there is at least fair evidence that the intervention improves important health outcomes. But the GDT concludes that the balance of the benefits, harms, and costs is too close to justify a general recommendation.</p>
<p>Recommendation: D</p> <p>Language:^a Recommendation against routinely providing the intervention to eligible patients.</p> <p>Evidence: The GDT found at least fair evidence that the intervention is ineffective, or that harms or costs outweigh benefits.</p>
<p>Recommendation: I</p> <p>Language:^a The evidence is insufficient to recommend for or against routinely providing the intervention. (At the discretion of the GDT, the recommendation may use the language "option," but must list all the equivalent options.)</p> <p>Evidence: Evidence that the intervention is effective is lacking, of poor quality, or conflicting and the balance of benefits, harms, and costs cannot be determined.</p>
Consensus-Based Recommendations¹
<p>Language:^a The language of the recommendation is at the discretion of the GDT, subject to approval by the Guideline Directors Group.</p> <p>Evidence: The level of evidence must be specified as "Good," "Fair," or "Insufficient" to match the language regarding evidence-based recommendations above. However, do not use the A, B, C, D, I labels which are only intended to be used for evidence-based recommendations.</p>

^a All statements specify the population for which the recommendation is intended.

Note that most consensus-based recommendations will have evidence grade "Insufficient." For the rare consensus-based recommendations which have "Good" or "Fair" evidence, the evidence must support a different recommendation, because if the evidence were good or fair, the recommendation would usually be evidence based. In this kind of consensus-based recommendation the evidence label should point this out, eg, "Good, supporting a different recommendation."

is sufficient evidence of good to fair quality, an evidence-based recommendation can be made. It is important that the language of the recommendation accurately reflects the strength of the supporting evidence (Table 1). In addition to the strength of the evidence, other factors such as the magnitude of benefit and potential or actual harms also need to be taken into consideration. In most cases, the process of weighing benefits and risks requires some degree of subjective judgment. What distinguishes an evidence-based methodology is the transparency of the rationale where all assumptions and value judgments are made explicit.

The actual guidance to the provider or other user of a guideline is called a recommendation.

The recommendation is the actionable statement, driven by the evidence summary, that tells the provider what treatment, test, etc, should be provided to the patient. For national KP guidelines, the language of the recommendation is strictly defined for consistency and clarity (Table 3).

Discussion of Hypothetical Case Study by Workshop Participants: "Memoryboost for Dementia"

The case study was based on the KP national dementia guideline, updated in 2004.³ The problem Formulation, Evidence Search and Evidence Summary Tables were taken from the guideline with minor changes. The assignment for the small groups in this track was to develop a Recommendation based on the material from this guideline and the information from the Common Methodology, Tables 1 and 2. The groups did not have time to

create the supporting Rationale for their Recommendation, so the authors developed one based on the contents of the discussion of one of the small groups.

Problem Formulation

In the case of "Memoryboost" the specific "PICO" clinical question was, "In men and women with dementia, does "Memoryboost" as compared to placebo or other drugs used to treat dementia result in improved cognitive and functional ability?"

Evidence Search

The databases searched included the Cochrane Database of Systematic Reviews, PubMed and others. The type of studies specified included systematic reviews, meta-analyses and randomized clinical trials. The search terms, driven by the "PICO" criteria, included the diagnoses of Alzheimer's disease and dementia and the same pharmacologic agents listed above under "I" (intervention) and "C" (comparison). The search identified over 200 systematic reviews and clinical trials, of which 16 met the inclusion criteria of the PICO and were of sufficient quality to be summarized for the team.

Evidence Summary

For the case study, Table 2, slightly simplified from the original, summarizes the results for one of the five studies about "Memoryboost" that were included in the evidence summary. The five tables were reviewed by the discussion groups. All groups were concerned about the quality of these articles. The primary issues were bias (all five were funded by the manufacturer of "Memoryboost"), duration of follow-up (three were only for 24 weeks; the other two were for one year)

and the clinical significance of the results (difference in MMSE between groups <2 points on the 30 point MMSE scale in which 2-3 points is considered to be clinically significant). Information about adverse effects, which were relatively frequent, but mild to moderate in severity, was also presented to these groups. Finally, after their initial deliberations and conclusions, the groups were told that the cost of the drug is about \$3 per day.

Recommendation

While the five small groups believed that overall, the evidence did not support a clear recommendation to use the drug in the treatment of dementia, they came to three different conclusions about the quality of the evidence and the recommendation that was driven by that evidence. All used the definitions from the Common Methodology in stating their conclusions (Table 3).

Some groups chose a “C” recommendation: “No recommendation for or against routine provision of “Memoryboost” for the management of cognitive and functional decline in mild to moderate dementia.”

Recommendation: C

Language: No recommendation for or against routine provision of the intervention. (At the discretion of the guideline development team, the recommendation may use the language “option,” but must list all the equivalent options.)

Evidence: Evidence is sufficient to determine the benefits, harms, and costs of an intervention, and there is at least fair evidence that the intervention improves important health outcomes. But the guideline development team concludes that the balance of the benefits, harms, and costs is too close to justify a general recommendation.

Some groups chose a “D” recommendation: “Memoryboost” is not recommended for the management of cognitive and functional decline in mild to moderate dementia.”

Recommendation: D

Language: Recommendation against routinely providing the intervention to eligible patients.

Evidence: The guideline development team found at least fair evidence that the intervention is ineffective, or that harms or costs outweigh benefits.

Some groups chose an “I” recommendation: “Memoryboost is an option for the management of cognitive and functional decline in patients with mild to moderate dementia.”

Recommendation: I

Language: The evidence is insufficient to recommend for or against routinely providing the intervention. (At the discretion of the guideline development team, the recommendation may use the language “option,” but must list all the equivalent options.)

Evidence: Evidence that the intervention is effective is lacking, of poor quality, or conflicting and the balance of benefits, harms, and costs cannot be determined.

The fact that these groups came up with three different interpretations of the evidence may be surprising to some. However, close inspection of the wording of these three recommendations reveals that when the evidence is weak or conflicting, reasonable people may disagree about whether it is sufficient to make a recommendation and whether it demonstrates net benefit or harm. In fact, most of the controversies in the development and approval of national KP guidelines occur in situations where the evidence is weak or inconsistent.

Rationale

The rationale for the “C” recommendation: “No recommendation for or against the routine provision of “Memoryboost” for the management of cognitive and functional decline in patients with mild to moderate dementia,” might go as follows. “There are five RCTs that evaluate “Memoryboost” compared with placebo. The GDT is concerned about the quality of all of them. The primary issues are bias (all five were funded by the manufacturer of “Memoryboost”), duration of follow-up (only 24 weeks for three studies and one year for the other two) and the clinical significance of the results. (The difference in MMSE between groups is <2 points on the 30 point MMSE scale in which 2-3 points is considered to be clinically significant.) Adverse effects were relatively frequent, but mild to moderate in severity. Based on the lack of a clinically significant difference in outcomes, combined with concerns about bias and duration of follow-up, the GDT concludes that the evidence is insufficient to make a recommendation.”

This example clearly shows how the rationale statement explicitly links the recommendation to the underlying evidence.

Discussion

National KP guidelines are created under the aegis of the KP National Guideline Directors by development teams that represent all of the appropriate specialties, stakeholder departments, as well as methodologic experts. These teams use only the highest-quality evidence, critically appraised and interpreted in light of the clinical experience of the team members. This process produces high-quality, evidence-based guidelines for implementation across the Pro-

gram. In several cases, eg, coronary artery disease, the guidelines are implemented by scores of care managers and others across the Program, using high-quality regional disease registries. As a result, some KP regions are giving some of the best care in the nation, which is reflected in the Healthplan, Employer Data and Information Set (HEDIS).

In the future, improvements in methodology by other organizations may result in guidelines of sufficient quality that KP will be able to focus on adapting those guidelines to the specifics of the KP care delivery system and benefits.

Finally, each KP Region has developed many guidelines by other methods. In general they represent

the consensus of the providers in that region. Many, if not most, of these guidelines are consistent with the medical literature, but without a comprehensive, well-documented evidence-based process, they are not likely to be adopted by other regions. The current national guidelines development process can serve as the mechanism to facilitate national guidelines on specified conditions once their importance reaches national significance and the published literature is sufficient to support an evidence-based process. ❖

References

1. Kaiser Permanente. The Guideline Directors Group. A common methodology and process for

interregional guidelines [monograph on the Intranet]. 3rd ed. 2004 [cited 2005 Mar 22]. Available from: <http://cl.kp.org/pkc/national/iknow/irgsg/CommonMethodology.doc>.

2. Evidence-Based Medicine Working Group; G Guyatt, D Rennie, editors. Users' guides to the medical literature; a manual for evidence-based clinical practice. Chicago: AMA; 2002.
3. Kaiser Permanente. Care Management Institute. CMI dementia guidelines [monograph on the Intranet]. [Oakland (CA)]: Care Management Institute; 2002 [revised 2004 Jan; cited 2005 Mar 22]. Available from: http://cl.kp.org/pkc/national/cmi/programs/dementia/guideline/files/dementia_guidelines_2004.pdf.

Beauty

When I am working on a problem
I never think about beauty.
I only think about how to solve the problem.
But when I have finished,
if the solution is not beautiful,
I know it is wrong.

— Buckminster Fuller, 1895-1983, engineer, designer, and architect