Study Guide
Session 1
Introduction to the Course:
Introduction to Team-Based Learning
Introduction to the Concept of Relevance of Medical Information

Allen F. Shaughnessy, PharmD, MMedEd
Professor of Family Medicine

The aims of this session are to:

1) Help you understand the course structure and process, including responsibilities and grading process;
2) Practice, without grading, the team-based learning approach (see below) that we will use;
3) Introduce the concept of "usefulness" as it pertains to medical information and the distinction between patient-oriented evidence and disease-oriented evidence.

Specific Objectives: By completing the initial reading and participating in class, students should be able to:

1) Describe the process of preparing for and participating in class activities;
2) List how their preparation and participation will affect their class grade;
3) Explain why not all information is created equal as a way of honing information searching;
4) Distinguish between patient-oriented and disease-oriented evidence

This study guide provides an outline of the concepts necessary to meet these objectives. It contains hyperlinks to short videos, web pages, or articles that explain the concepts in other ways or in greater detail. You can follow these hyperlinks if the explanations and examples I've given you are not sufficient to help you understand and to help you prepare to participate in class.

Preparation for Session 1

Before class time: Read this study guide

Welcome!

This course on evidence-based medicine is designed to build on the information you learned in the epidemiology and biostatistics course in the fall and prepare you for the Introduction to Clinical Reasoning and other courses coming up (as well as the rest of your career). We'll be looking at medical research from a completely different viewpoint in this course -- as *consumers* (i.e. users) of research and not as *producers* of research (researchers). You will learn completely new skills and knowledge that build on what was covered in epi/biostats. The "cognitive load" (amount of information) is low and many of the concepts are relatively easy; it's the application of these skills and this information that's harder. For this reason will be using a "team-based learning" approach in this course.

Why study Evidence-Based Medicine?

You learned about study design, biases inherent in conducting research, and how to design the "perfect study" in epidemiology and biostatistics. However, in clinical medicine, we can't wait for publication of a perfect study before making decisions regarding the care of patients. Instead we have to use the results of imperfect studies conducted by humans on humans.

The methods we will discuss will identify problems in the studies that render their results not applicable to the patients we see in practice, either because the results do not apply to our patient populations (they lack external validity) or because the study design and conduct have flaws that make us question the results (lack of internal validity).

In this course we will discuss how to evaluate four types of medical information: 1) original research evaluating new therapies; 2) original research evaluating new diagnostic tests; 3) reviews that summarize original research or synthesize new conclusions from original research; and, 4) practice guidelines that are offered to guide practice.

Evaluating research in the ways we will discuss is not an academic exercise but is designed to answer the central question in medical practice:

> *Am I practicing in a way that is most likely to help my patients live longer, better, or both?*

Our goal is not to decide whether research is "good" or "bad." Instead, we will be asking whether the evidence we have available to us is *useful*, helping us to meet our goal of improving the lives of people. There is no greater mission in medicine.

Introduction to Team-Based Learning

Team-based learning is an approach to teaching and learning that involves completing assigned pre-work followed by completing graded, individual tests *before* each class. During class we will work in small groups (similar to your PBL groups), in the Sackler lecture hall, completing the group tests and practicing application of the concepts covered in the prework.

This approach is designed to maximize your time in class by avoiding some of the problems with lectures (boredom), small group work (slackers), and test surprises (the final exam will consist of the exact same type of exercises completed in class).

Many of the concepts in the course are not difficult to understand, and there is little information to commit to memory. Application of the concepts, though, is harder. This approach to learning will help you master the *application* of the concepts, a skill that will support you through upcoming clinical work and through the rest of your career.

 **Watch TEAMLead at Duke-NUS (9:04)**

Each class will require preparation and completion of an (IRAT)"individual readiness assessment test." The study guide for each class will provide an explanation of the material with links to additional resources and examples. You may wish to do further learning on your own to prepare for the course if this information is not enough.

Introduction to the Course

Grading

	Percent of Grade
TUSK Individual Readiness Assessment Tests(5% for each test x 4)	**20**
In-Class Group IRAT Tests	**30**
Class Attendance	**10**
Final Exam	**30**
Homework Assignment	**10**

Some important aspects of team-based learning (TBL):

1) ***Class attendance to all sessions is mandatory*.** Class time will be used to work in groups to complete the same questions asked on the IRAT, and then apply the ideas to an example. *Classes will not be videorecorded or streamed*.

2) **All assigned preparation work *must be completed* before each session**. The syllabus contains an explanation of the concepts with hyperlinks to additional readings and videos that may be helpful. The syllabus can be printed out but it is designed to be read on a computer or tablet screen. If you understand the concepts by reading only the material in the syllabus, there is no need to follow the hyperlinks (*i.e., there is no "gotcha" information buried in one of the hyperlinked resources*). Expect to spend about 45 minutes in preparation before each class.

3) **The on-line Individual Readiness Assessment Tests (IRATs) *must be completed* before the start of class.** There will be one IRAT for each class period (n = 5). Note- only IRAT #2-5 will be graded. These will be available on TUSK. The IRATs will be opened for completion immediately after the previous class ends and will remain open until immediately before the class starts. This step assures me, and your teammates, that you come to the class prepared. It also allows me to see where the group, as a whole, has had difficulty. These are individual (no group work) and comprise 20 percent of your grade.

TUSK On-Line IRAT Test	Class Date	Open Date/Time	Close Date/Time	Test w/Answers Posted
#1- Introduction (not graded but required)	1-May	Fri., April 26 @12pm	Wed., May 1 @8am	Wed., May 1 @10am
#2- Research on Therapy	8-May	Wed., May 1 @12pm	Wed., May 8 @10am	Wed., May 8 @12pm
#3- Research on Testing	10-May	Wed., May 8 @12pm	Fri., May 10 @8am	Fri., May 10 @10am
#4- Reviews of Meta-Analysis	15-May	Fri., May 10 @12pm	Wed., May 15 @10am	Wed., May 15 @12pm
#5- Practice Guidelines	22-May	Wed., May 15 @12pm	Wednesday, May 22 @8am	Wednesday, May 22 @10am

Following class, the IRATs will be available again on TUSK, this time with the answers provided, for your use to review before the final exam.

4) **There will be four graded IRAT tests given in-class during the course.** These IRAT tests will be done with your assigned group and each member of the group will be assigned the same grade for the test. These in-class, group tests will account for 30 percent of your grade.

5) **Homework Assignment:** There will be one homework assignment for this course and it is to be done individually. The assignment will be due on **Friday, May 24, 2013 by 4pm** and should be submitted to Dr. Shaughnessy by e-mail at allen.shaughnessy@tufts.edu. The assignment is to critically assess an original research study, practice guideline, or review article that you identified for your library literature assignment. You will use one of the worksheets associated with the classes, determine the article's validity, and come to a conclusion regarding your original question. See appendix 2. This assignment will be worth 10 percent of your grade.

6) **You get credit for showing up!** Students with unexcused absences will lose 2 percentage points for each missed class and will not receive credit for the group's team score for that session**.**

7) **Except for the last session, there is no need for computers during class, and group work in class will not use computers.** The questions and exercises will be answered using the combined knowledge and wisdom of the group.

8) **Appeals:** If you strongly feel that an answer on your group work is also correct, you may submit an appeal, by team, explaining why the question is poorly worded or, based on appropriate references, why the answer you selected is also correct. Only teams that submit accepted appeals will have their grades corrected on both the IRAT and the group work. See appendix 1 for a complete explanation.

Frequently Asked Questions

Should we work on the IRATs as a group outside of class?
No. The individual tests should be completed on your own.

Will our group have to meet outside of class?
No. All group work will be done in class.

Will I need to bring a computer or pad to class?
A computer will be needed (one per group) only for the last class.

How to Succeed in the Evidence-Based Medicine Course:

1. Complete the pre-assigned preparation work, as indicated in the electronic syllabus, before each EBM session.
2. Complete the graded, on-line IRAT test on TUSK **before** each session (see chart above for the open and close date of each of these tests).
3. Attend each EBM TBL session and participate actively in group work assigned (attendance is mandatory and you will be graded on it).
4. Complete the graded, in-class IRAT group test for sessions 2-5.
5. Complete the homework assignment (as described above and in appendix 2) for the course and return it via e-mail to Dr. Shaughnessy at allen.shaughnessy@tufts.edu by **Friday, May 24, 2013** at 4pm.
6. Take the final exam on Friday, May 24, 2013 at 8:30am.

Introduction to the Concept of *Relevance* of Medical Information

Shock! Not all information in medicine is useful

What is "evidence"? Typically, we have a lot of information available to us when making decisions. Take, for example, the treatment of hypertension. There are over 356,000 articles in the Medline database! Of these,

- 276,421 involve human subjects
- 28,859 are clinical trials
- 52,338 are review articles
- In the last year, 72,028 articles were published

If you were to try to read all of the medical information published each month, 24 hours a day, 7 days a week, within 1 week you would already be two years behind.

In addition, there are 508 clinical practice guidelines* just on the treatment of patients with hypertension.

Even if you tried to read only the clinical trials involving humans published in the last year, you would have to read 2.73 studies every day.

Fortunately, we don't have to read all these articles. since of this information is better than other types of information.

Concept: Usefulness of information

How do we begin to sort through information? This equation helps us conceptualize the usefulness of information available to us:

$$\textbf{Usefulness of any information}^{\dagger} = \frac{\textit{Relevance} \times \textit{Validity}}{\textit{Work}}$$

The usefulness of any information for decision-making varies with how relevant and how valid it is, and how much work (time, effort, money) it takes to find that information.

To define relevance,‡ we consider two different types of outcomes. The information available to us can present either *patient-oriented* or d*isease-oriented* evidence.

---

* National Guideline Clearinghouse

† Shaughnessy AF, Slawson DC, Bennett JH. Becoming an Information Master: A Guidebook to the Medical Information Jungle. The Journal of Family Practice 1994;39(5):489-99.

‡ Validity will be considered in upcoming sessions. Work is addressed next year.

Patient-Oriented vs. Disease-Oriented Evidence

Patient-oriented evidence is direct information that tells us that a medical intervention affects morbidity, mortality, or quality of life. In other words, patient-oriented evidence is that which tells patients that what we are going to do for them, or ask them to do, will make them live longer or live better.

Disease-oriented evidence on the other hand, considers the pathology, physiology, pharmacology and the etiology of disease. For example, an effect of a treatment on a blood test or on a risk factor is disease-oriented evidence. These are surrogates rather than the disease. For example, no one ever dies of high blood cholesterol; they die due to the *effect* of high cholesterol.

The distinction between patient-oriented and disease-oriented evidence is crucial, since there are many examples, in the history of medical care, where effects on disease did not translate into improvements in patient-oriented outcomes. The Table gives some examples of when disease-oriented evidence and patient-oriented evidence conflict:

Disease-Oriented Outcome	Patient-Oriented Outcome
Intensive treatment can lower blood glucose levels in patients with type 2 diabetes	Intensive treatment in patients with type 2 diabetes does not decrease mortality.
Beta-carotene and vitamin E are good antioxidants	Neither beta-carotene or vitamin prevents cardiovascular disease or cancer
The drug varenicline can help smokers stop smoking (which should lead to a decrease in cardiovascular events)	Varenicline increases the risk of cardiovascular events
Medications can decrease irregular heartbeats in patients with asymptomatic arrhythmias	Medical treatment of asymptomatic arrhythmias *increases* mortality by **10%**[§]

Watch "The Surrogate Battle: Is Lower Always Better? (4:05)

Optional reading: Becoming an Information Master

[§] An estimated 65,000 people died because of the once-common practice of treating asymptomatic arrhythmias based on disease-oriented evidence. That's more people than are listed on the Vietnam War Memorial. See: *Moore TJ. Excess Mortality Estimates. Deadly Medicine: Why Tens of Thousands of Heart Patients Died in America's Worst Drug Disaster. New York, NY: Simon & Schuster, 1995; 281–9.*

Appendix 1.

Readiness Assessment Test Appeals Instructions

Purposes of the appeals process:

1. Clarify uncertainty about your understanding of the concepts.

2. Give additional recognition and credit when "missing" a question was caused by:

    - Ambiguity in the reading material.
    - Disagreement between the reading material and our choice of the "correct" answer.
    - Ambiguity in the wording of the question.

Guidelines for preparing successful appeals:

Appeals are granted when they demonstrate that you understood the concept(s) but missed the question anyway or that your confusion was due to ambiguity in the reading material. As a result:

- If the appeal is based on ambiguity in the question, you should:

    1. Identify the source of ambiguity in the question and,
    2. Offer an alternative wording that would have helped you to avoid the problem. The appeal will be accepted if I decide to use the question in the future.

- If the appeal is based on either inadequacies in the reading material or disagreement with our answer, you should:

    1. State the reason(s) for disagreeing with our answer and,
    2. Provide specific references from the reading material to support your point of view. If I agree with your logic, I will award credit.

Impact if appeals on test scores:

When an appeal is accepted on a question that a <u>group</u> has missed (no <u>individual</u> appeals will be accepted):

1. It "counts" i.e., the points missed will be added to:

    - Their group score
    - The score of any individual in the group who answered the same way the group did
    - Only those groups that appeal.

Group member(s) who had the original correct answer will continue to receive credit on the question.

Appendix 2. Homework Assignment

Who: Each person will submit the assignment. This is not a group assignment.

What: You will critically appraise the research paper, practice guideline, or review article you identified to answer the question posed in the library search assignment. If you did not use one of these types of paper for your assignment, you may substitute another research paper, guideline or review article.

When: The homework should be submitted no later than the end of the course, **Friday, May 24, 2013 at 4pm.**

Where: E-mail a copy of the paper and the evaluation form to allen.shaughnessy@tufts.edu.

How: For your identified article, answer the questions on the appropriate worksheet (see each syllabus section). For each question, circle "yes" or "no" and give a brief (1-2 sentence) explanation supporting your choice. At the bottom of the worksheet, answer the following question:

Based on this exercise, explain whether your conclusion has changed to the original question, "After considering the evidence you found, what conclusions and decisions have you made concerning your patient?" This answer should be two or three sentences.

Grading rubric: This assignment must be completed and contributes 10% to the overall grade. The assignment will be scored as follows:

70%: Correct choices for the relevance and validity questions and supporting explanations. The percentage will be evenly divided among the questions (the number of questions differs among the worksheets).

30%: Appropriate conclusion and explanation regarding the question, "After considering the evidence you found, what conclusions and decisions have you made concerning your patient?"

Evidence-Based Medicine

Spring 2013 Schedule

Date	Session Title	Time	Location	Assignment Due
Wed. May 1	Introduction to EBM (M)	8:15-9:05am	Sackler	TUSK IRAT # 1 Test-Introduction
Wed. May 8	Research on Therapy(M)	10:00-12:00pm	Sackler	TUSK IRAT #2 Test-Research on Therapy
Fri. May 10	Research on Testing(M)	8:00-10:00am	Sackler	TUSK IRAT #3 Test-Research on Testing
Wed. May 15	Reviews of Meta-Analysis (M)	10:00-12:00pm	Sackler	TUSK IRAT #4 Test-Reviews of Meta-Analysis
Wed. May 22	Practice Guidelines (M)	8:00-10:00am	Sackler	TUSK IRAT #5 Test-Practice Guidelines
Fri. May 24	Final Exam	11:00-12:30pm	See Posted Assign.	Homework Assignment

Attendance is MANDATORY at all sessions in the EBM course.

****See table in syllabus for open and close dates/times of the above assignments.***

Study Guide
Session 1
Introduction to the Course:
Introduction to Team-Based Learning
Introduction to the Concept of Relevance of Medical Information

Allen F. Shaughnessy, PharmD, MMedEd
Professor of Family Medicine

The aims of this session are to:

1) Help you understand the course structure and process, including responsibilities and grading process;
2) Practice, without grading, the team-based learning approach (see below) that we will use;
3) Introduce the concept of "usefulness" as it pertains to medical information and the distinction between patient-oriented evidence and disease-oriented evidence.

Specific Objectives: By completing the initial reading and participating in class, students should be able to:

1) Describe the process of preparing for and participating in class activities;
2) List how their preparation and participation will affect their class grade;
3) Explain why not all information is created equal as a way of honing information searching;
4) Distinguish between patient-oriented and disease-oriented evidence

This study guide provides an outline of the concepts necessary to meet these objectives. It contains hyperlinks to short videos, web pages, or articles that explain the concepts in other ways or in greater detail. You can follow these hyperlinks if the explanations and examples I've given you are not sufficient to help you understand and to help you prepare to participate in class.

Preparation for Session 1

Before class time: Read this study guide

Welcome!

This course on evidence-based medicine is designed to build on the information you learned in the epidemiology and biostatistics course in the fall and prepare you for the Introduction to Clinical Reasoning and other courses coming up (as well as the rest of your career). We'll be looking at medical research from a completely different viewpoint in this course -- as *consumers* (i.e. users) of research and not as *producers* of research (researchers). You will learn completely new skills and knowledge that build on what was covered in epi/biostats. The "cognitive load" (amount of information) is low and many of the concepts are relatively easy; it's the application of these skills and this information that's harder. For this reason will be using a "team-based learning" approach in this course.

Why study Evidence-Based Medicine?

You learned about study design, biases inherent in conducting research, and how to design the "perfect study" in epidemiology and biostatistics. However, in clinical medicine, we can't wait for publication of a perfect study before making decisions regarding the care of patients. Instead we have to use the results of imperfect studies conducted by humans on humans.

The methods we will discuss will identify problems in the studies that render their results not applicable to the patients we see in practice, either because the results do not apply to our patient populations (they lack external validity) or because the study design and conduct have flaws that make us question the results (lack of internal validity).

In this course we will discuss how to evaluate four types of medical information: 1) original research evaluating new therapies; 2) original research evaluating new diagnostic tests; 3) reviews that summarize original research or synthesize new conclusions from original research; and, 4) practice guidelines that are offered to guide practice.

Evaluating research in the ways we will discuss is not an academic exercise but is designed to answer the central question in medical practice:

> *Am I practicing in a way that is most likely to help my patients live longer, better, or both?*

Our goal is not to decide whether research is "good" or "bad." Instead, we will be asking whether the evidence we have available to us is *useful*, helping us to meet our goal of improving the lives of people. There is no greater mission in medicine.

Introduction to Team-Based Learning

Team-based learning is an approach to teaching and learning that involves completing assigned pre-work followed by completing graded, individual tests *before* each class. During class we will work in small groups (similar to your PBL groups), in the Sackler lecture hall, completing the group tests and practicing application of the concepts covered in the prework.

This approach is designed to maximize your time in class by avoiding some of the problems with lectures (boredom), small group work (slackers), and test surprises (the final exam will consist of the exact same type of exercises completed in class).

Many of the concepts in the course are not difficult to understand, and there is little information to commit to memory. Application of the concepts, though, is harder. This approach to learning will help you master the *application* of the concepts, a skill that will support you through upcoming clinical work and through the rest of your career.

 **Watch TEAMLead at Duke-NUS (9:04)**

Each class will require preparation and completion of an (IRAT)"individual readiness assessment test." The study guide for each class will provide an explanation of the material with links to additional resources and examples. You may wish to do further learning on your own to prepare for the course if this information is not enough.

Introduction to the Course

Grading

	Percent of Grade
TUSK Individual Readiness Assessment Tests(5% for each test x 4)	**20**
In-Class Group IRAT Tests	**30**
Class Attendance	**10**
Final Exam	**30**
Homework Assignment	**10**

Some important aspects of team-based learning (TBL):

1) ***Class attendance to all sessions is mandatory*.** Class time will be used to work in groups to complete the same questions asked on the IRAT, and then apply the ideas to an example. *Classes will not be videorecorded or streamed*.

2) **All assigned preparation work *must be completed* before each session**. The syllabus contains an explanation of the concepts with hyperlinks to additional readings and videos that may be helpful. The syllabus can be printed out but it is designed to be read on a computer or tablet screen. If you understand the concepts by reading only the material in the syllabus, there is no need to follow the hyperlinks (*i.e., there is no "gotcha" information buried in one of the hyperlinked resources*). Expect to spend about 45 minutes in preparation before each class.

3) **The on-line Individual Readiness Assessment Tests (IRATs) *must be completed* before the start of class.** There will be one IRAT for each class period (n = 5). Note- only IRAT #2-5 will be graded. These will be available on TUSK. The IRATs will be opened for completion immediately after the previous class ends and will remain open until immediately before the class starts. This step assures me, and your teammates, that you come to the class prepared. It also allows me to see where the group, as a whole, has had difficulty. These are individual (no group work) and comprise 20 percent of your grade.

TUSK On-Line IRAT Test	Class Date	Open Date/Time	Close Date/Time	Test w/Answers Posted
#1- Introduction (not graded but required)	1-May	Fri., April 26 @12pm	Wed., May 1 @8am	Wed., May 1 @10am
#2- Research on Therapy	8-May	Wed., May 1 @12pm	Wed., May 8 @10am	Wed., May 8 @12pm
#3- Research on Testing	10-May	Wed., May 8 @12pm	Fri., May 10 @8am	Fri., May 10 @10am
#4- Reviews of Meta-Analysis	15-May	Fri., May 10 @12pm	Wed., May 15 @10am	Wed., May 15 @12pm
#5- Practice Guidelines	22-May	Wed., May 15 @12pm	Wednesday, May 22 @8am	Wednesday, May 22 @10am

Following class, the IRATs will be available again on TUSK, this time with the answers provided, for your use to review before the final exam.

4) **There will be four graded IRAT tests given in-class during the course.** These IRAT tests will be done with your assigned group and each member of the group will be assigned the same grade for the test. These in-class, group tests will account for 30 percent of your grade.

5) **Homework Assignment:** There will be one homework assignment for this course and it is to be done individually. The assignment will be due on **Friday, May 24, 2013 by 4pm** and should be submitted to Dr. Shaughnessy by e-mail at allen.shaughnessy@tufts.edu. The assignment is to critically assess an original research study, practice guideline, or review article that you identified for your library literature assignment. You will use one of the worksheets associated with the classes, determine the article's validity, and come to a conclusion regarding your original question. See appendix 2. This assignment will be worth 10 percent of your grade.

6) **You get credit for showing up!** Students with unexcused absences will lose 2 percentage points for each missed class and will not receive credit for the group's team score for that session**.**

7) **Except for the last session, there is no need for computers during class, and group work in class will not use computers.** The questions and exercises will be answered using the combined knowledge and wisdom of the group.

8) **Appeals:** If you strongly feel that an answer on your group work is also correct, you may submit an appeal, by team, explaining why the question is poorly worded or, based on appropriate references, why the answer you selected is also correct. Only teams that submit accepted appeals will have their grades corrected on both the IRAT and the group work. See appendix 1 for a complete explanation.

Frequently Asked Questions

Should we work on the IRATs as a group outside of class?
No. The individual tests should be completed on your own.

Will our group have to meet outside of class?
No. All group work will be done in class.

Will I need to bring a computer or pad to class?
A computer will be needed (one per group) only for the last class.

How to Succeed in the Evidence-Based Medicine Course:

1. Complete the pre-assigned preparation work, as indicated in the electronic syllabus, before each EBM session.
2. Complete the graded, on-line IRAT test on TUSK **before** each session (see chart above for the open and close date of each of these tests).
3. Attend each EBM TBL session and participate actively in group work assigned (attendance is mandatory and you will be graded on it).
4. Complete the graded, in-class IRAT group test for sessions 2-5.
5. Complete the homework assignment (as described above and in appendix 2) for the course and return it via e-mail to Dr. Shaughnessy at allen.shaughnessy@tufts.edu by **Friday, May 24, 2013** at 4pm.
6. Take the final exam on Friday, May 24, 2013 at 8:30am.

Introduction to the Concept of *Relevance* of Medical Information

Shock! Not all information in medicine is useful

What is "evidence"? Typically, we have a lot of information available to us when making decisions. Take, for example, the treatment of hypertension. There are over 356,000 articles in the Medline database! Of these,

- 276,421 involve human subjects
- 28,859 are clinical trials
- 52,338 are review articles
- In the last year, 72,028 articles were published

If you were to try to read all of the medical information published each month, 24 hours a day, 7 days a week, within 1 week you would already be two years behind.

In addition, there are 508 clinical practice guidelines* just on the treatment of patients with hypertension.

Even if you tried to read only the clinical trials involving humans published in the last year, you would have to read 2.73 studies every day.

Fortunately, we don't have to read all these articles. since of this information is better than other types of information.

Concept: Usefulness of information

How do we begin to sort through information? This equation helps us conceptualize the usefulness of information available to us:

$$\textbf{Usefulness of any information}^{\dagger} = \frac{\textit{Relevance x Validity}}{\textit{Work}}$$

The usefulness of any information for decision-making varies with how relevant and how valid it is, and how much work (time, effort, money) it takes to find that information.

To define relevance,‡ we consider two different types of outcomes. The information available to us can present either *patient-oriented* or d*isease-oriented* evidence.

---

* National Guideline Clearinghouse

† Shaughnessy AF, Slawson DC, Bennett JH. Becoming an Information Master: A Guidebook to the Medical Information Jungle. The Journal of Family Practice 1994;39(5):489-99.

‡ Validity will be considered in upcoming sessions. Work is addressed next year.

Patient-Oriented vs. Disease-Oriented Evidence

Patient-oriented evidence is direct information that tells us that a medical intervention affects morbidity, mortality, or quality of life. In other words, patient-oriented evidence is that which tells patients that what we are going to do for them, or ask them to do, will make them live longer or live better.

Disease-oriented evidence on the other hand, considers the pathology, physiology, pharmacology and the etiology of disease. For example, an effect of a treatment on a blood test or on a risk factor is disease-oriented evidence. These are surrogates rather than the disease. For example, no one ever dies of high blood cholesterol; they die due to the *effect* of high cholesterol.

The distinction between patient-oriented and disease-oriented evidence is crucial, since there are many examples, in the history of medical care, where effects on disease did not translate into improvements in patient-oriented outcomes. The Table gives some examples of when disease-oriented evidence and patient-oriented evidence conflict:

Disease-Oriented Outcome	Patient-Oriented Outcome
Intensive treatment can lower blood glucose levels in patients with type 2 diabetes	Intensive treatment in patients with type 2 diabetes does not decrease mortality.
Beta-carotene and vitamin E are good antioxidants	Neither beta-carotene or vitamin prevents cardiovascular disease or cancer
The drug varenicline can help smokers stop smoking (which should lead to a decrease in cardiovascular events)	Varenicline increases the risk of cardiovascular events
Medications can decrease irregular heartbeats in patients with asymptomatic arrhythmias	Medical treatment of asymptomatic arrhythmias *increases* mortality by **10%**[§]

Watch "The Surrogate Battle: Is Lower Always Better? (4:05)

Optional reading: Becoming an Information Master

[§] An estimated 65,000 people died because of the once-common practice of treating asymptomatic arrhythmias based on disease-oriented evidence. That's more people than are listed on the Vietnam War Memorial. See: *Moore TJ. Excess Mortality Estimates. Deadly Medicine: Why Tens of Thousands of Heart Patients Died in America's Worst Drug Disaster. New York, NY: Simon & Schuster, 1995; 281–9.*

Appendix 1.

Readiness Assessment Test Appeals Instructions

Purposes of the appeals process:

1. Clarify uncertainty about your understanding of the concepts.

2. Give additional recognition and credit when "missing" a question was caused by:

    - Ambiguity in the reading material.
    - Disagreement between the reading material and our choice of the "correct" answer.
    - Ambiguity in the wording of the question.

Guidelines for preparing successful appeals:

Appeals are granted when they demonstrate that you understood the concept(s) but missed the question anyway or that your confusion was due to ambiguity in the reading material. As a result:

- If the appeal is based on ambiguity in the question, you should:

    1. Identify the source of ambiguity in the question and,
    2. Offer an alternative wording that would have helped you to avoid the problem. The appeal will be accepted if I decide to use the question in the future.

- If the appeal is based on either inadequacies in the reading material or disagreement with our answer, you should:

    1. State the reason(s) for disagreeing with our answer and,
    2. Provide specific references from the reading material to support your point of view. If I agree with your logic, I will award credit.

Impact if appeals on test scores:

When an appeal is accepted on a question that a <u>group</u> has missed (no <u>individual</u> appeals will be accepted):

1. It "counts" i.e., the points missed will be added to:

    - Their group score
    - The score of any individual in the group who answered the same way the group did
    - Only those groups that appeal.

Group member(s) who had the original correct answer will continue to receive credit on the question.

Appendix 2. Homework Assignment

Who: Each person will submit the assignment. This is not a group assignment.

What: You will critically appraise the research paper, practice guideline, or review article you identified to answer the question posed in the library search assignment. If you did not use one of these types of paper for your assignment, you may substitute another research paper, guideline or review article.

When: The homework should be submitted no later than the end of the course, **Friday, May 24, 2013 at 4pm.**

Where: E-mail a copy of the paper and the evaluation form to allen.shaughnessy@tufts.edu.

How: For your identified article, answer the questions on the appropriate worksheet (see each syllabus section). For each question, circle "yes" or "no" and give a brief (1-2 sentence) explanation supporting your choice. At the bottom of the worksheet, answer the following question:

Based on this exercise, explain whether your conclusion has changed to the original question, "After considering the evidence you found, what conclusions and decisions have you made concerning your patient?" This answer should be two or three sentences.

Grading rubric: This assignment must be completed and contributes 10% to the overall grade. The assignment will be scored as follows:

70%: Correct choices for the relevance and validity questions and supporting explanations. The percentage will be evenly divided among the questions (the number of questions differs among the worksheets).

30%: Appropriate conclusion and explanation regarding the question, "After considering the evidence you found, what conclusions and decisions have you made concerning your patient?"

Study Guide
Session 2
Determining the Validity of Research on a Therapy
Allen F. Shaughnessy, PharmD, MMedEd

The aims of this session are to:

1) Build on your previous study of epidemiology and biostatistics to apply the concepts to medical research used in clinical medicine
2) Introduce you to some additional issues of study validity
3) Practice how to quickly and accurately evaluate research methods and findings about treatments used in clinical medicine

Specific Objectives: By completing the initial reading and participating in class, students should be able to:

1) List and describe the major threats to validity in studies evaluating treatments
2) Interpret the results of clinical studies
3) Use a set of questions to quickly evaluate a study for validity and relevance.

This study guide provides an outline of the concepts necessary to meet these objectives. It contains hyperlinks to short videos, web pages, or articles that explain the concepts in other ways or in greater detail. You can follow these hyperlinks if the explanations and examples I've given you are not sufficient to help you understand and to help you complete the readiness assessment test.

What is evidence-based medicine, and why is it important?

Evidence based medicine, according to the innovators of this approach, is "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research."[*]

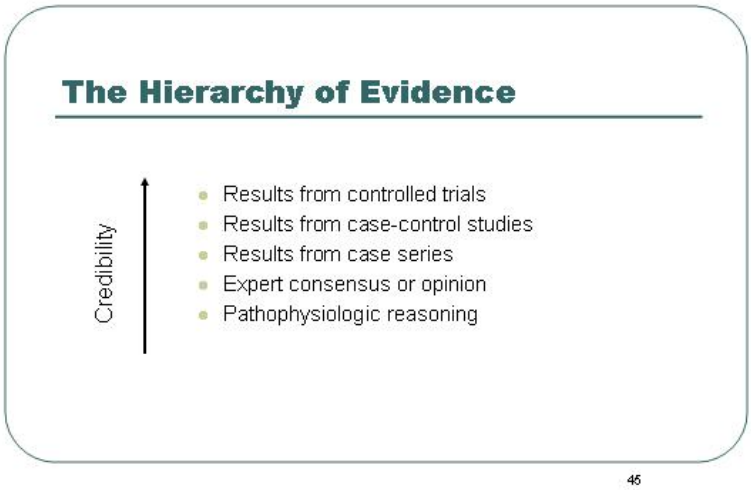
Further, another good explanation is that, "Evidence based medicine (EBM) has not developed a new concept of evidence; its major contribution lies in the emphasis it places on a hierarchy of evidential reliability, in which conclusions related to evidence from controlled experiments are accorded greater credibility than conclusions grounded in other sorts of evidence."[†]

For a third definition, closely in line with these, watch: EBM defined (4:27)

[*] http://www.bmj.com/content/312/7023/71
[†] http://www.bmj.com/content/329/7473/1024.

So, the idea behind evidence-based medicine is that there is a **hierarchy of evidence**, with some types of evidence being more likely to represent the truth than others are.

This idea, in turn, is based on the concept that truth is a probability rather than an absolute in medicine. A treatment may increase the likelihood of certain patients will receive benefit, but we are by no means certain that everyone will benefit.



For example, whenever we study a new treatment, there are various ways that a study could be designed. We might take an epidemiologic approach. Or, we might study the pharmacology of a drug. We might try the treatment on one person and publish a case report, or study a bunch of people in a case series. The credibility of the results depends on the study design.

Issues of study validity

The following questions can be used to determine whether a study is sufficiently valid. There is a worksheet that can be used to keep track of the answers.

Read: "Evaluating and Understanding Articles About Treatment" (3 pages)

Are the studied patients similar enough to your patients that you can apply the results in your practice?

Another aspect of a study is whether the **population studied is similar** to your population. The severity and likelihood of illness and the response to treatment will vary based on where study subjects are found.

Example

What is the risk of a non-febrile seizure in child at some point following a febrile seizure?

This graph show the difference in studies reporting the risk of febrile seizures in children who have had a seizure as a result of a high fever. Studies conducted in pediatric neurologist offices report a follow-up seizure rate of about 40% - 60%. However, studies in emergency departments report a follow-up seizure rate of about 5% or less. Clearly, children who end up seeing a pediatric neurologist are quite different from those simply seen in an emergency department.

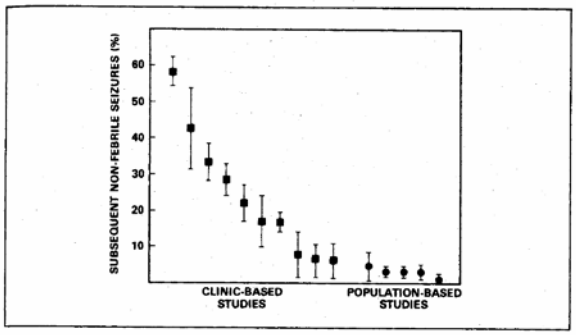


Figure 6-1. The risk of nonfebrile seizures following febrile convulsions. ■ = studies from specialty clinics and hospitals; ● = studies from general populations. (Adapted from J. H. Ellenberg and K. B. Nelson. Sample selection and the natural history of disease. Studies of febrile seizures. *J.A.M.A.* 243:1337, 1980.)

Were the subjects randomly assigned?

When it comes down to accurately evaluating any therapeutic intervention (drug, procedure, or even whether to obtain a particular diagnostic test), we generally should turn, if we can, to **a randomized control trial**.

Randomization is really the best protection that we have against being misled. In comparative studies, in which biases such as the compliance effect and the placebo effect are controlled, randomized studies are less likely than other types of studies to show that a treatment was effective when it really isn't. A longer explanation of why this is so is in the YouTube video:

YouTube **Why is randomization important? (5:03)**

Example

How well does warfarin work to prevent deaths in patients who have had a heart attack?

Thomas Chalmers and colleagues compared the differences in reported rates of anticoagulation with warfarin in patients with an acute heart attack. They found that the demonstrated effectiveness dropped quite a bit when comparing results from randomized controlled studies with results from historical control (before-after) studies.

The value of randomization

- 32 controlled trials of anticoagulation for the treatment of acute myocardial infarction
- Results by type of study:

	Relative Risk Reduction	Case fatality rate
Historical control	**42%**	**38.3%**
Controlled trial	**33%**	**29.2%**
Randomized controlled trial	**31%**	**19.6%**

A 26% drop in reported effectiveness

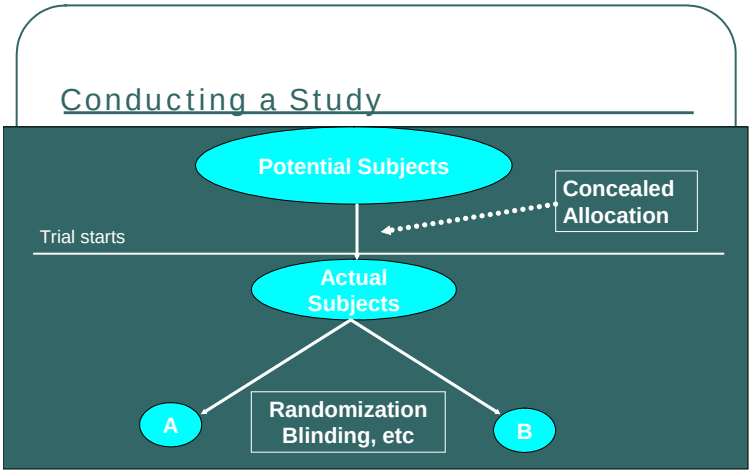
Chalmers TC, et al. N Engl J Med 1977;297:1091-6.

58

Were steps taken to conceal the treatment assignment from study personnel entering patients into the study?

The gold standard for a therapeutic trial is no longer simply just the concept of randomization, but is also whether or not **allocation assignment was concealed** from the enrolling investigator ("concealed allocation"). The definition of concealed allocation: Did the investigators know to which group a potential subject would be assigned before they were actually enrolled in this study? Concealed allocation is used in randomized studies to protect the integrity of the randomization process. Randomization is meant to prevent selection bias, and if there was no concealed allocation, the study is again susceptible to selection bias. In this case, the selection bias is occurring during the enrollment period and causes the study population to be unrepresentative of the population of potential subjects.

Trials that do not use conceal allocation consistently overestimate the benefit of the treatment by 40%.[‡] The recent changes in recommendations for breast cancer screening using mammography were prompted by a discovery almost 15 years ago that the lack of allocation concealment biased studies that evaluated the effectiveness of screening mammography.



Allocation concealment is not the same thing as blinding. Allocation concealment occurs before a study begins, during the process of selecting patients for a study. It is possible to have a study that is blinded, but does not conceal allocation. It's also possible to have a non blinded study that does conceal allocation.

Watch "What is concealed allocation?" (1:51)

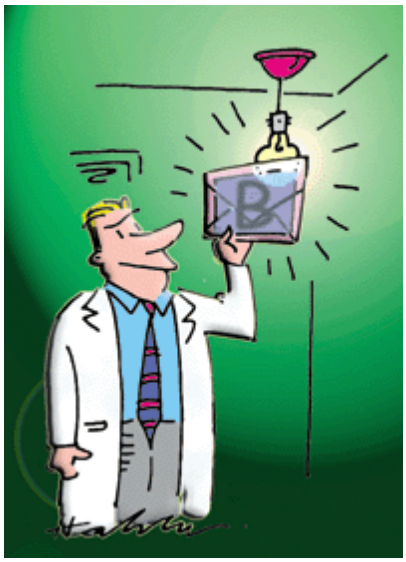
***Read:* Screening Mammography: Controversies and Headlines**

[‡] JAMA. 1995 Feb 1;273(5):408-12.

Read: Allocation concealment for examples of blinded studies without allocation concealment as well as non-blinded studies with concealed allocation.

Example

For example, a study performed in the early 1990s, before allocation concealment was considered an important issue, evaluated the benefit of artificial surfactant for newborn premature infants in the neonatal intensive care unit. It was possible for physicians and nurses caring for infants to hold the study envelopes up to a light and determine whether or not the next baby that could potentially be enrolled in the study would either receive surfactant or placebo.



Here's a possible scenario: Let's say the investigators held study envelopes up to the light to determine whether the child would be put in the surfactant or placebo group. If they were to be put in the placebo group, children who were marginal and likely not to survive, regardless of any intervention, might have been enrolled. Children with a reasonable chance for survival may not have been enrolled in the study (but simply given surfactant outside the study).

As a result, sicker children could have been selectively enrolled into the placebo group. This selective enrollment may have greatly overestimated the benefit of surfactant because the more healthy children with the higher likelihood of survival were not enrolled in the control arm of the study. This study was published in the medical literature as a randomized, double-blind controlled trial. Since the authors did not specify that the envelopes were sealed and opaque, readers would not know that allocation was concealed.

Were all patients who entered the trial properly accounted for at its conclusion?

We want to be able to follow the history of every subject to know what happened to them. We want to avoid selective analysis of the data, which rarely happens in published work today. Most journals require a figure with a flow diagram (below) showing what happened to every patient. What percentage of patients were lost to follow-up? When more than 10 - 20% of patients are lost to follow-up it is difficult to accept the results of the study.

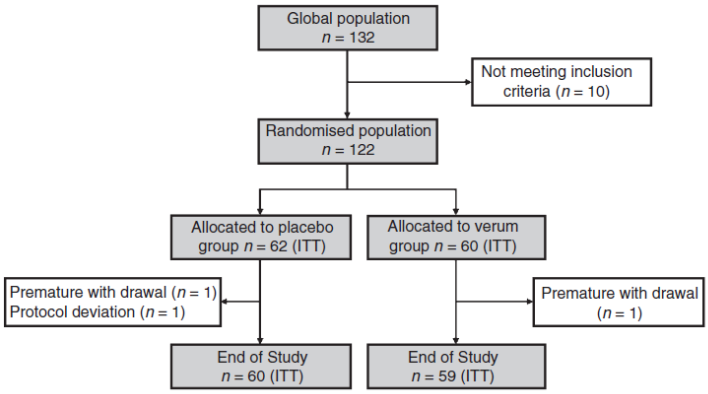


Figure 2 | Diagram of study flow. Verum, *B. bifidum* MIMBb75.

Were patients analyzed in the groups to which they were randomized ("intention-to-treat" analysis)?

Data should be analyzed by **intention-to-treat**, meaning that patients were analyzed in the group to which they were initially assigned regardless of whether they actually received the treatment. This method of analysis more accurately reflects the real world effect of an intervention where not all patients are compliant with treatment. Studies reporting results only from subjects in the intervention group proven to have taken the medicine ("on-treatment" analysis) compared with the placebo group are actually comparing compliant individuals with both compliant and non-compliant ones. Regardless of what intervention is studied, compliant patients will usually do better than those who are non-compliant, making the treatment look more effective than it actually is.

Watch "What is intention-to-treat analysis?" (1:11)

Watch "Intention-to-treat analysis: What is it and why is it important?" (4:43)

A review of useful statistics[§]

After we've decided a study is valid we need to **understand the results**. Understanding just a few statistics is all that's necessary. Briefly, here are some statistics that are useful:

P value: "P" stands for the *probability* that the difference between two averages, rates, etc – whatever we're measuring – is due to chance. It is the likelihood that the difference we see in the numbers is not "real", in that it represents an actual difference due to one treatment over another, but that, if we did the study again, we might find the difference to no longer be present. So, a P-value of 0.05 tells us that we have a 5% risk that the difference is due to chance, or a 95% likelihood that the difference we see represents a real difference.

Watch What is a P-Value? (5:56)

Self-test: Which of the results on this Table are not due to chance?

Table 3| **Comparison of Cohen-Mansfield agitation inventory (CMAI) total score between control and intervention (stepwise protocol for treatment of pain) groups using repeated measures analysis of covariance (ANCOVA)***

	Mean (SD) CMAI total score		Effect of intervention on CMAI total†		Intracluster correlation coefficient‡
Week	Control group	Intervention group	Estimate (95% CI)	P value	
0	56.2 (16.1), n=177	56.5 (15.2), n=175	—	—	0.162
2	53.9 (17.0), n=161	52.0 (19.5), n=158	−3.6 (−0.5 to −6.7)	0.022	0.261
4	52.5 (16.3), n=160	49.4 (19.0), n=148	−4.1 (−0.9 to −7.4)	0.012	0.231
8	52.8 (16.8), n=157	46.9 (18.7), n=147	−7.0 (−3.7 to −10.3)	<0.001	0.226
12	52.5 (16.0), n=152	50.3 (20.3), n=142	−3.2 (0.1 to −6.4)	0.058	0.253

*Baseline score as covariate and least squares weighted by number of patients within cluster; P value from multivariate test of intervention was 0.002, and cross effect between week and intervention was <0.001.
†Variable estimate by week of effect of intervention on CMAI score from estimated model.
‡Proportion of total variance between clusters, and measured within framework of ANCOVA.

Answer here

[§] To read more, see: http://tinyurl.com/cx68dxx

Number needed to treat (NNT): The NNT tells us how many people we need to treat instead of not treat, or the number that need to be treated with one therapy instead of another, for one *additional* person to benefit. It takes into account the idea that some people will benefit even without therapy, some people will benefit from another treatment, and it accounts for the fact that treatment usually is not 100% effective and thus some people who receive treatment will not benefit.

It is calculated based only on results that are presented as rates (ie, in percents). We calculate by taking the different between two rates (the rates of cure, the rate of death, etc) and divide that number into 100.

$$\text{NNT} = \frac{100}{\text{\% in treatment group - \% in control group}}$$

(use the absolute value; that is, don't worry about a minus sign)

Watch "What is number needed to treat?" (0:40)

Watch The NNT Tutorial (3:57)

Self-test: From the study results below, calculate how many additional patients would need to be treated with antibiotic rather than placebo for one additional person not to develop new lesions.

> Results: On 30-day follow-up (successful in 69% of patients), we observed fewer new lesions in the antibiotic (4/46; 9%) versus placebo (14/50; 28%) groups, difference 19%, 95% confidence interval 4% to 34%, *P*<.02.

Answer here

Relative risk: Relative risk helps us understand the difference between two rates – rates of death, rates of a side effect, etc. Depending on how the results are worded, it can represent the risk of harm or the risk of benefit. For example, a relative risk of 1.3 tells us that the likelihood of something happening is 30% higher in one group vs. the other; a relative risk of 0.7 tells us the likelihood of something happening is 30% lower in one group vs. the other. If the relative risk = 1.0, then there is no difference.

Unfortunately, the relative risk doesn't take into account the baseline risk, or the risk of no treatment, so we have to ask ourselves, 30% of what?

 **Relative risk (7:56)**

The **confidence interval** tells us the range of possible results. A 95 percent confidence interval (95% CI) indicates that if the study were repeated 100 times, the study results would fall within this interval 95 times. In the example above for number needed to treat, the rate difference was 19% (95% CI = 4%-34%). The 95% confidence interval tells us that, if we performed the study again many times, we would find a rate difference somewhere between 4% and 34% 95 out of 100 times.

 **Watch "What are Confidence Intervals? (1:29)**

 **Watch Interpretation of Confidence Interval (1:54)**

*The **power** of a study is its ability to find a difference between the groups if a group truly exists.* We only need to concern ourselves with power if there was not a difference between treatments – then we need to ask whether the study had sufficient power.

 **Watch Power of a Study (2:03)**

Using a worksheet to evaluate articles for relevance and validity

The goal of using this or other worksheets is to quickly determine whether it is worth taking the time to read a research study (or a synopsis of that study). It allows determination, based on answers to the questions, whether the information is relevant to you, and whether the study design has sufficient rigor to apply the results to your patients.

The questions focus on the study design issues described above. The first 6 questions address study "musts" – they ask about issues of relevance or validity that must be present if the study results are to be applied to clinical practice. The answers to these questions must be yes regardless of the answers to the rest of the questions.

The goal of this worksheet is ***not*** to determine whether a study is "good" or "bad." Instead, we will use it to determine whether the results reported by the study are likely to occur if we use the same approach in our patients. As a result, the information is either useful to us, or not.

Answers to the self-test questions:

Self-test: Which of the results on this Table are not due to chance?

> All but week 12 are statistically significant.

Back to the self-test

Self-test: From the study results below, calculate how many additional patients would need to be treated with antibiotic rather than placebo for one additional person not to develop new lesions.

> NNT = 100/(28% - 9%) = 5.26. One additional person would not have follow-up lesions after 30 days for every 6 patients treated with antibiotic instead of placebo.

Back to the self-test

Return to instructions

A Worksheet for Articles about Treatment

Determine *Relevance*

Is this article worth taking the time to read? If the answer to any of these questions is No, it may be better to read other articles first.

Based on the conclusion of the abstract:

A. Did the authors study an outcome that patients would *care* about? (Be careful to avoid results that require extrapolation to an outcome that truly matters to patients)

Yes (go on) No **(stop)**

B. Is the problem studied one that is *common* to your practice and the intervention feasible?

Yes (go on) No **(stop)**

C. Will this information, if true, require you to *change* your current practice?

Yes (go on) No **(stop)**

Determine *Validity*

If the answers to all three questions above are Yes, then continued assessment of the article is mandatory.

D. Population

1. Are the studied patients similar enough to your patients that you can apply the results in your practice? Yes No **(Stop)**

E. Study design

1. Was it a controlled trial? Yes No **(Stop)**
2. Were the subjects randomly assigned? Yes No **(Stop)**
3. Were steps taken to conceal the treatment assignment from study personnel entering patients into the study? Yes No
4. Were patients, providers and outcome assessors "blind" to treatment? Yes No

F. Study conduct

1. Were all patients who entered the trial properly accounted for at its conclusion?
   a. Was follow-up complete? Yes No
   b. Were patients analyzed in the groups to which they were randomized ("intention-to-treat" analysis)? Yes No
2. Were the intervention and control groups similar? (Table 1) Yes No

G. Study results

1. What were the results? _____

_____

2. Are the results clinically as well as statistically significant? Yes No
3. If a negative trial, was the power of the study adequate? Yes No
4. Were there other factors that might have affected the outcome? Yes No
5. How will it change your practice?

_____

_____

Return to instructions

Study Guide
Session 3
Determining the Value of a Diagnostic Test

Allen F. Shaughnessy, PharmD, MMedEd
Allen.Shaughnessy@tufts.edu

The aims of this session are to:

1. Review and expand upon some ideas of disease screening and testing that have been covered in the epidemiology;
2. Introduce the concept of the difference between the technical precision and the clinical precision of a screening or diagnostic test;
3. Practice how to quickly and accurately evaluate studies evaluating a new diagnostic or screening test to determine their validity

Specific Objectives: By completing the initial reading and participating in class, students should be able to:

1. Explain the different ways the same testing procedure can be used;
2. Use the test characteristics of sensitivity, specificity, predictive values, and likelihood ratios to correctly interpret test results
3. Explain how pre-test probability can affect test accuracy
4. Determine whether the results of a study evaluating a test are relevant and valid and support that test's use in clinical practice.
5. Explain why the fact that simply using a test because it is highly sensitive or specific may not always be beneficial.

This study guide provides an outline of the concepts necessary to meet these objectives. It contains hyperlinks to short videos, web pages, and articles that explain the concepts in other ways or in greater detail. You can follow these hyperlinks if the explanations and examples I've given you are not sufficient to help you understand and to help you complete the readiness assessment test.

Purposes of testing

A test, whether a physical exam maneuver, a laboratory test, an imaging study, or a performance measure (such as reading an eye chart), can play various roles, depending on our needs. The same test can:

1) **Screen** for an unapparent disease in an asymptomatic individual (*screening*)
2) **Identify** a disease in a patient suspected of having that disease (*case finding*)
3) **Confirm** or **refute** results of another test (*confirmation*)
4) **Evaluate** the effectiveness of treatment (*monitoring*)
5) Allow estimation of **prognosis** to help guide treatment decisions (*prognosis*)

Example

Same test, different uses

Measuring someone's blood pressure can be used for various purposes, depending on the setting and the need.

1) Screening: Checking blood pressure at a health fair or at a self-testing station – asymptomatic patients with low likelihood of hypertension;
2) Case finding: Checking patients' blood pressure at the start of an office visit – patients are still at low risk of hypertension (*unless already diagnosed*), but are at higher risk than when screened (as in example #1), since they are seeking health care
3) Confirmation: Checking someone's blood pressure who has had a previously high reading via screening or in the office;
4) Monitoring: Checking blood pressure in someone with diagnosed hypertension; and,
5) Prognosis: Checking blood pressure in patients with acute myocardial infarction or stroke can be used to estimate 30-day mortality.

So What? Bayes' Theorem

Tests are not perfect – all tests have some risk of false positive and false negative results.* These false results will vary based on the likelihood of the test being positive or negative *before we ever do the test*. This likelihood is called *prior probability*, *pre-test probability,* or *prevalence* of disease.

Simply put, Bayes' theorem is:

post-test probability = Pre-test probability, given the test result

The effect of pre-test probability on test results is intuitive: if someone is highly unlikely to have a disease but a test is positive, it makes sense to think that the test is wrong (i.e. still has a low post-test probability). Conversely, if another person has dramatic symptoms and signs of a disease but the test comes back negative, it makes sense that the test is a false negative (i.e. still has a high post-test probability).

For more information, see: Bayes' Theorem and Breast Cancer (9:56)

* Except, perhaps, the "birth test" and the "death test"

For example,

- In a 70-year-old male with sudden onset chest pain that worsens with exercise, the likelihood of myocardial infarction, before additional testing, is 63%. A negative electrocardiogram would be **falsely negative** in 1-in-6 men like this patient.
- In a 40-year-old male with sudden onset chest pain that worsens with exercise, the likelihood of myocardial infarction, before additional testing, is around 0.6%.[†] A positive electrocardiogram would be **falsely positive** in 11 of 12 men like this patient.[‡]

Pretest probability is determined a number of ways—sometimes, we have done epidemiologic studies to find the population prevalence of certain diseases. Sometimes, the pretest probability is a best-guess estimate because we don't always have that kind of data on every condition.

So, when screening for disease, the pre-test probability of whatever we are screening for will be low, making positive test results suspect. Case-finding increases the probability somewhat, and the factors that prompt testing to confirm a diagnosis increase probability even more. *The key point is that the same test will perform differently given this background probability.*

We can quantify how likely a test is to be falsely positive using simple calculations. What is *not* intuitive for most people, though, is how likely a test is likely to be false in situations of low pre-test likelihood.

Example

Some examples of high false positives

1. In the U.S., lyme disease prevalence varies by geography, highest in Connecticut and lowest in Texas. False positive results:
   - Connecticut (20% prevalence): 17% false positive
   - Texas (2% prevalence): 72% false positive

2. The prevalence of breast cancer in women increases with age. For a 30-year-old woman undergoing a screening mammography:
   - Prevalence: 1 in 235 (0.43%)
   - False positive rate: 94%

[†] Ruling out coronary artery disease in primary care: development and validation of a simple prediction rule.
[‡] Acute chest pain in the emergency room. Identification and examination of low-risk patients.

Why is this important?

1. Hazards of testing

Patients who have a positive test result, even if it is found later to be falsely positive, can have lasting psychological results:

Three years after a false-positive mammogram result, women are more anxious about having breast cancer and have prolonged psychosocial effects such as, "my sense of well-being is less" and "my relationship with other people is worse"

In a study of infants diagnosed with jaundice (for which the evidence is inconclusive as to whether they benefit from diagnosis and treatment):

- Mothers are more likely to completely stop breast feeding
- Mothers were more likely to have never left their baby alone with anyone else, including the father
- The infant had more office visits and emergency department visits

This result has been called the "vulnerable child syndrome" that occurs, in this case, as a result of testing that may not produce benefit.

2. Lack of benefit of testing

Physicians commonly express the belief that patients want diagnostic testing to check for serious but unlikely illnesses. This type of testing is often ordered with the aim of reassuring patients. However, "Diagnostic tests for symptoms with a low risk of serious illness do little to reassure patients, decrease their anxiety, or resolve their symptoms, although the tests may reduce further primary care visits."

JAMA Intern Med. 2013;173(6):407-416.

Technical vs. clinical precision of a test

Sensitivity and *specificity* of a test are characteristics used to judge the intrinsic technical precision of a test. They, for the most part, are insensitive to changes in prevalence.

Positive predictive value and *negative predictive value* are characteristics used to judge the clinical performance of a test, i.e., how well does the test represent the truth in clinical practice? Predictive values *are* sensitive to prevalence. As illustrated above, the same test, with the same sensitivity and specificity, will result in different rates of false positive and false negative results, depending on the pre-test likelihood of a positive or negative result.

In the above examples, the technical precision of the tests look pretty good:

Lyme disease detection
Sensitivity: 95%
Specificity: 95%

Breast cancer detection (mammography)
Sensitivity: 79%
Specificity: 89%

However, the effect of prevalence makes the tests perform much worse, in clinical practice, than is intuitively obvious.

Lyme disease positive predictive value
High prevalence: 83%
Low prevalence: 28%

Mammography positive predictive value
High prevalence: 34%
Low prevalence: 0.8%

> In other words, in a place where Lyme disease is highly prevalent, 83% of the time, when the test is positive, the patient actually has the disease. Another way to interpret this is to say that in a high prevalence area, there is a 17% false-positive rate. In a low prevalence area, when the test is positive, 28% of those patients will truly have Lyme disease.
>
> As a result, even though the sensitivity and specificity of these tests are high, the tests will be falsely positive a majority of the time when the prevalence of disease is low. In the case of Lyme disease, about 3 out of 4 people told they have Lyme disease will not; and 92 out of 100 low likelihood women with a finding on mammogram will not have breast cancer.

Watch Sensitivity and Specificity – getting a feel (8:14)

Calculating Sensitivity, Specificity, and Predictive Values

There are several ways to calculate test characteristics, though test characteristics typically are calculated using a 2x2 table, listing the true status of the disease at the top and the test results along the side. The key to setting up the table is **to list the disease at the top of the square**.

	Disease truly present	**Disease truly absent**
Positive test	True Positives (TP) a	False Positives (FP) b
Negative test	c False Negatives (FN)	d True Negatives (TN)

Sensitivity

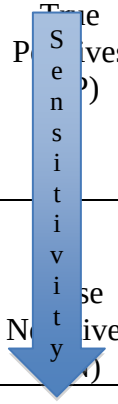
- Is the percent of patients *with the disease* who have a *positive test*
- *= TP/(TP + FN)*
- *= a/(a+c)*

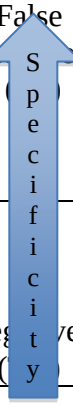
Specificity

- Is the percent of patients without the disease who have a negative test
- = TN/(TN + FP)
- = d/(d + b)

Using the graph, the calculations proceed “down” and “up” the columns:

	Disease truly present	**Disease truly absent**
Positive test	True Positives (TP) a	False Positives (FP) b
Negative test	c False Negatives (FN)	d True Negatives (TN)





Positive predictive value

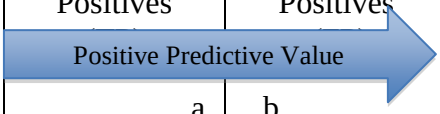
- Is the percent of patients *with a positive test* who *have the disease*
- = TP/(TP + FP)
- = a/(a + b)

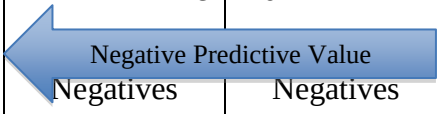
Negative predictive value

- Is the percent of patients with a *negative test* who *do not have the disease*
- *= TN/(TN + FN)*
- *= d/(d + c)*

Using the graph, the calculations proceed “left” and “right” across the columns:

	Disease truly present	**Disease truly absent**
Positive test	True Positives a	False Positives b
Negative test	c False Negatives (FN)	d True Negatives (TN)





Watch “[Sensitivity, specificity, and predictive values]” (2:34)

Practice:

Here are the results of a test for HIV.

	HIV present	**HIV absent**
HIV test +	475	4975
HIV test -	25	94525

Calculate the test characteristics:

Sensitivity: Positive Predictive Value:

Specificity: Negative Predictive Value:

Answers here

Practice: Sample USMLE Step 1 question:

To protect blood supplies from contamination, screening for all donors for hepatitis C is required. The screening test has a sensitivity of 95% and a specificity of 90% and is used on a sample of donors in which 10% are known to have hepatitis C infection. Which of the following is the best estimate of the chance that a donor who tests negative is actually free of infection?

A. 45%
B. 50%
C. 85%
D. 90%
E. 95%
F. 99%

Calculations shown at: Calculating NPV from sensitivity and specificity (4:11)

Using test characteristics: SnNout and SpPin

Tests that are highly sensitive are very good at identifying patients with disease. As a result of this quality, we can be sure that *negative* tests are truly negative. This relationship can be remembered using the mnemonic *SnNout*:

SnNout: If a test is highly ***S*ensitive**, and the test result is ***N*egative**, we can rule **out** the disease

Conversely, tests that are highly specific are negative unless patients truly have the disease. As a result, we can be sure that positive tests are truly positive. This relationship can be remembered using the mnemonic *SpPin*:

SpPin: If a test is highly ***S*pecific**, and the test result is ***P*ositive**, we can rule **in** the disease.

This rule holds when the likelihood is not too low (e.g., screening) or too high (e.g., confirmatory testing).

Watch Using SpPin and SnNout (2:23)

Using test characteristics: Predictive values

The **good** thing about predictive values *is that they vary with prevalence of disease*. By knowing the prevalence, we can calculate the predictive value of a test for individual situations, helping us to make decisions regarding treatment or further testing.

The **bad** thing about predictive values is the same: *they vary with prevalence of disease*. It is often hard to calculate the pre-test likelihood of disease, especially in the moment, and we find it hard to find a resource that lists prevalence. Fortunately, there are calculators that can provide estimates of prevalence based on physical findings and calculate predictive values based on the test's sensitivity and specificity.

Here is an example of a clinical calculator that helps determine the pre-test probability of patients having sore throat based on symptoms.

Combining all of the above: Likelihood ratios

The likelihood ratio gives us an understanding of how strongly a test result helps us rule in or rule out a disease. A **positive likelihood ratio** compares the likelihood that someone with the disease in question has a positive test as compared with someone who doesn't have the disease. A negative likelihood **ratio** compares the likelihood that someone without the disease in question will have a negative test as compared with someone who does have the disease.

Likelihood ratios are calculated based on sensitivity and specificity of the test:

Positive Likelihood Ratio = sensitivity / (1-specificity)
Negative Likelihood Ratio = (1-sensitivity) / specificity

In practice, likelihood ratios (LRs) are used in two ways.

1) To calculate post-test odds of a disease, given pre-test odds and the LR:

Pretest odds of disease X LR = Post-test odds of disease

Most of us, however, don't think in terms of odds, which makes the calculations difficult. However, many calculators are available that will take pre-test probability, convert it to pre-test odds, calculate the post-test odds from the LR, and convert this odds back into a probability. There are also paper or computer-based nomograms that will make the calculations easy.

2) To give a general interpretation of a test's quality. The size of a LR helps us to understand how valid a test result might be. General rules:

Likelihood ratio	Interpretation
>10	Good test to rule-in disease with a positive result
5 - 10	Moderately able to rule-in with a positive test
2 - 5	Small increase in probability with a positive test
1-2	No change in probability with a positive test
0.5 – 1	No change in probability with a negative test
0.2 – 0.5	Small increase in probability with a negative test
0.1 – 0.2	Moderately able to rule-out disease with a negative test
< 0.1	Good test to rule-out disease with a negative result

These are good, general rules, for gauging the quality of a diagnostic test. However, because prevalence still impacts a test's results, these rules don't apply if the pre-test probability is very high or very low.

For example, a test with a positive LR of 500 will only have a positive predictive value of 50% given a pre-test probability of 1 in 5,000.

Beyond Test Characteristics: What Makes a Test Truly Helpful?

As we've discussed, tests can identify unknown disease, confirm presumed disease, help with estimates of prognosis, or monitor response to treatment. But can a test be held responsible to do more?

Limits of Testing:

Screening: Early[4] identification of disease is only beneficial if treating before symptoms occur results in greater benefit than treating based on symptoms.

Watch Overdiagnosed. (it's funny and informative, but long; over an hour)

Diagnostic Testing: A diagnosis, though the *sina qua non* of medicine, is, at its essence only a label placed on a patient that is useful when selecting the right treatment. Therefore, testing leading to a diagnosis is **only** helpful if it leads to a change in *treatment.*

A Hierarchy of Evidence Regarding Tests[5]

As a result of these limitations, it's not enough to simply say that a test has a good sensitivity and specificity (or predictive values). We need our tests to do more:

Basic Criteria: Is the test sufficiently sensitive and specific?

We have many tests with low sensitivity and specificity.

Minimally useful: The test changes diagnosis

As a result of a positive test, we now have a label to put on a patient. That doesn't mean that we've done anything other than categorize their set of signs and symptoms. What's better is to. . .

More useful: The test changes treatment

A test that results in changes in treatment is a good start. However, tests don't always lead to changes. Sputum samples are often suggested in guidelines of the treatment of pneumonia. However, research has shown physicians frequently do not change treatment when the culture results are known a day or two later.

Very useful: The test changes outcomes

Routine monitoring (with A1c) did not affect outcomes in the United Kingdom Prospective Diabetes Study.[6]

---

[4] That is, before symptoms appear.

[5] *Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11:88-94*

[6] *Turner RC, et al. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 1998;352:837-53.*

On the other hand, testing in the emergency department to determine whether the patient does or doesn't have heart failure has been shown to decrease admissions, decrease length of stay, and speed the initiation of appropriate therapy.[7]

Maximum benefit: The test is worthwhile to patients and/or society.

Screening newborns for congenital hypothyroidism, phenylketonuria, and other diseases (but not all) results in early treatment that permanently alters the life of these children. On the other hand, screening for other diseases, and treating them (kernicterus, oxygenation) has not shown to be beneficial.

McCormack J, Greenhalgh T. Seeing what you want to see in randomised controlled trials: versions and perversions of UKPDS data. BMJ 2000;320:1720-1723)

[7] *Mueller C, Scholer A, Laule-Kilian K, et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. N Engl J Med 2004; 350: 647-54*).

(jump to worksheet)

Evaluating a Study about a New Diagnostic Test

Is the new test reasonable? What are its limitations?

The first step in evaluating a new screening test, diagnostic test, or other diagnostic maneuver is to determine whether the test could be used in clinical practice. A test that is too expensive, takes too long to perform, requires too much blood, or other limitation may not be worth considering.

Example
Now there is a rapid test for influenza, but in years past an influenza test had to be sent to a central laboratory. Results were returned in 5 to 7 days and were not useful for making decisions regarding treatment. Influenza testing was only useful after the fact for tracking patterns.

Is the reference (gold) standard appropriate?

A new test has to be compared to an existing standard. This reference standard, or "gold standard," should be the best currently available way of identifying disease. Lesser standards, with their own limitations, can confound the results.

Example
Some years ago, I conducted a study to determine whether two methods of skin caliper measurement and bioimpedence measurement were accurate measures of body fat; we wanted to determine the best way to identify the minimum weight for high school wrestlers. The gold standard for body fat analysis is total body immersion in water; we didn't have a facility to do this, which dramatically limited our ability to draw conclusions regarding the tests' accuracy.

Did all participants receive both the new test and the reference test?

All study subjects should receive both the new test and the gold standard test. Sometimes testing is done so that only patients with a positive result on the new test get the gold standard test, or vice versa. This approach biases the study.

Were the results of the test interpreted without knowledge (blinded) of the reference test result and vice versa?

We want to assure that both tests are interpreted independently, that is, without knowledge of the results of the other test. This approach prevents interpretation bias.

Example
Imagine checking the blood pressure in a patient just after someone else has. If you knew the first reading, you would expect your reading to be similar, and you might try a little harder to get a very similar. If, on the other hand, you don't know the results of the first blood pressure measurement, you will not have an inherent bias toward that result.

Were the patients enrolled randomly or consecutively?

Ideally, all patients eligible for testing would be enrolled; if not, we would like to see a random selection of patient. This criterion helps to assure a broad spectrum of patients, some with relatively mild disease and others with more severe disease, which gives us a more accurate understanding of the test's characteristics.

Example
Investigators conducted an early study of a test to determine whether patients with shortness of breath presenting to an emergency department have heart failure. The study was conducted during daylight hours (because that is when the research associate was available). However, patients with heart failure often don't develop symptoms until the evening when they are lying flat in bed. This study, therefore, likely enrolled patients whose disease was more severe, since they had symptoms during the day, rather than enrolling patients with a wide spectrum of heart failure.

Does the study population generalize to your practice?

Most studies will present the prevalence of disease for their population. The source of patients will also produce different spectra of disease states. Patients presenting to a primary care physician, for example, are likely to have less severe disease than patients who are subsequently referred to subspecialists. The study population should be similar to the population of patients you treat.

Jump to instructions

A Worksheet for Articles about Diagnostic Tests

Description of the tests:

1. Is the new test **reasonable**? What are its **limitations**? *(stop)*

2. Is the **reference (gold) standard** appropriate? YES (if yes, describe) NO *(stop)*

   EXPLAIN:

3. Did all participants receive **both** the new test and the reference test? YES NO *(stop)*

4. Were the results of the test interpreted without knowledge **(blinded)** of the reference test result and vice versa? YES NO

Study Population:

1. Were the patients enrolled randomly or consecutively? YES NO
2. Does the study population **generalize** to your practice? YES NO
   (Consider the spectrum of patient characteristics, co-morbidities, and clinical presentation)
   EXPLAIN:

D. *Test Characteristics:*

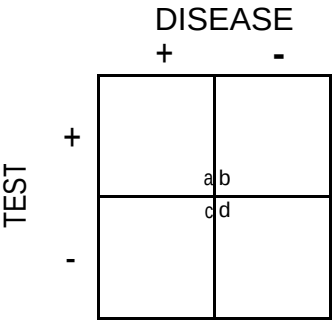
1. What are the **sensitivity**, **specificity** and **predictive values** of the test?

a. Sensitivity= $\frac{a}{a+c}$ _____

b. Specificity= $\frac{d}{b+d}$ _____

c. P.P.V.= $\frac{a}{a+b}$ _____

d. N.P.V.= $\frac{d}{d+c}$ _____



2. Calculate the **prevelance** of disease in the study $\frac{a+c}{a+b+c+d}$

3. How does this compare to your practice? _____

Revision 2003: Information Mastery Working Group. Adapted from material developed at McMaster University.

Answer to the HIV calculation

Sensitivity: 95% Positive Predictive Value: 8.7%

Specificity: 95% Negative Predictive Value: 99.9%

Back to the example

Study Guide
Session 4
Reading and Determining the Validity of Review Articles

Jessica Early, MD
JEarly@challiance.org

The aims of this session are to:

1. Help you identify review articles that are more likely to be valid; and,
2. Explain how to read the results of meta-analyses

Specific Objectives: By completing the initial reading and participating in class, students should be able to:

1. Differentiate synthesis and summary review articles
2. List the components of a systematic review
3. Use a worksheet to evaluate a systematic review for:
    a. The quality of the search and selection of evidence
    b. The quality of the included evidence
    c. Homogeneity of the results
    d. Evidence of publication bias
4. Interpret a forest plot of a meta-analysis

This study guide provides an outline of the concepts necessary to meet these objectives. It contains hyperlinks to short videos, web pages, or articles that explain the concepts in other ways or in greater detail. You can follow these hyperlinks if the explanations and examples I've given you are not sufficient to help you understand and to help you complete the readiness assessment test.

What are review articles?

Review articles summarize or analyze research previously published by others, rather than reporting new experimental results (although, as we will see, they also can report new data). They are often called "secondary literature" since they build on research literature, which is called "primary literature."

There are two main types of review articles. **Summary reviews** are the traditional type of review. They cover the full breadth of a particular topic, typically providing an overview of the disease etiology, diagnosis, prognosis, or management, and will usually address a number of questions. Experts usually write them. Book chapters are summary reviews. This type of review is useful and often fine for background questions.

 **Read Example of s summary review**

Synthesis reviews are systematic reviews, either with or without meta-analysis. Defining one or two specific questions, writers of systematic reviews carefully find all available evidence, evaluate its validity, and report their answer to the question. This type of review is useful when answering foreground questions.

 **Read Synthesis review example**

Meta-analysis is a statistical technique for combining the findings from independent studies. It can be performed following a systematic review, to treat the data from different studies as if they were from one large study, rather than simply counting the studies ("4 studies say it works, 2 studies say it doesn't, so I guess it works.")

 **Read Meta-analysis example**

 **Watch Meta-analysis (4:46)**

The Cochrane Collaboration

The Cochrane Collaboration is an international network of more than 28,000 people from over 100 countries (The New England Cochrane Center is housed at Tufts in the Center for Clinical Evidence Synthesis. The group has produced over 5,000 systematic reviews using a process that is considered to be the gold standard for systematic review and meta-analysis.

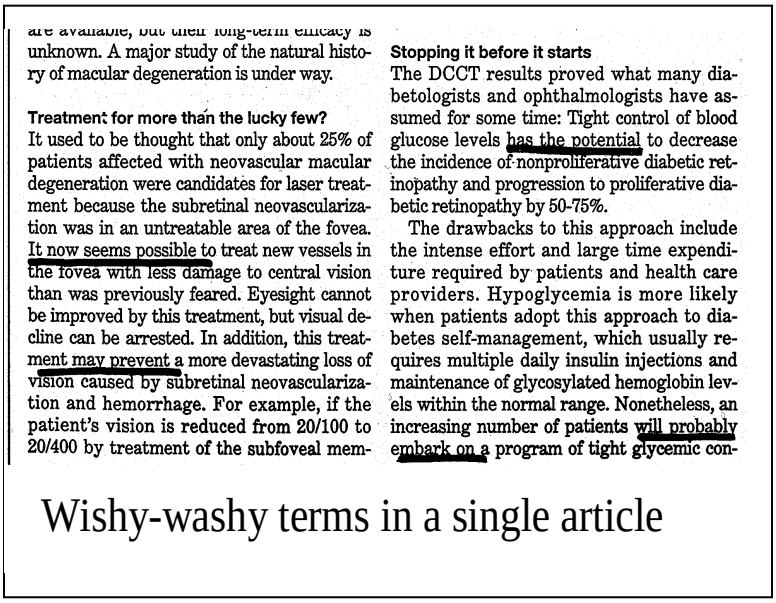
The logo for the Cochrane Collaboration is a forest plot (explained below) of the results of using corticosteroid treatment of pregnant women at risk of premature delivery. The use of this simple and inexpensive treatment decreases mortality in the newborns by 30% - 50%. It was not until publication of this meta-analysis in 1991 that maternity care physicians started using this treatment regularly, savings thousands of lives (more about the history of the Cochrane logo)

What are the issues with review articles?

Typically, authors of summary reviews are experts in the area of the review. As such, the review writer usually makes little or no attempt to be systematic in the formulation of the question they are addressing, searching for evidence, or summarizing the evidence they consider. As a result, the information in summary reviews has to be taken at face value.*

There are other issues with summary reviews:

1) *Misreferenced statements*. The citation at the end of the sentence doesn't support the statement. In some studies this has been as high as 40% of all citations.
2) *Information imposters*: The article seems to convey information but uses wish-washy phrase such as "may be effective" or "should be useful, leaving the reader unsure
3) *Missing information* due to a lack of a literature search.
4) *Lagging recommendations*. Recommendations in review articles may not be based on the best current evidence In an analysis of review articles of treatment of acute myocardial infarction, an average of 13 years passed between the time good evidence was available to support a treatment and the recommendation of that treatment for routine use in review articles. In review articles on the treatment of type 2 diabetes, most reviews did not accurately convey the results of a landmark study.
5) *Reliance on the expert's knowledge* rather than a systematic approach to evidence. The methodologic rigor of the review, in one study,† was inversely related to the self-rated clinical expertise of the review writer.



are available, but their long-term efficacy is unknown. A major study of the natural history of macular degeneration is under way.

Treatment for more than the lucky few?

It used to be thought that only about 25% of patients affected with neovascular macular degeneration were candidates for laser treatment because the subretinal neovascularization was in an untreatable area of the fovea. It now seems possible to treat new vessels in the fovea with less damage to central vision than was previously feared. Eyesight cannot be improved by this treatment, but visual decline can be arrested. In addition, this treatment may prevent a more devastating loss of vision caused by subretinal neovascularization and hemorrhage. For example, if the patient's vision is reduced from 20/100 to 20/400 by treatment of the subfoveal mem-

Stopping it before it starts

The DCCT results proved what many diabetologists and ophthalmologists have assumed for some time: Tight control of blood glucose levels has the potential to decrease the incidence of nonproliferative diabetic retinopathy and progression to proliferative diabetic retinopathy by 50-75%.

The drawbacks to this approach include the intense effort and large time expenditure required by patients and health care providers. Hypoglycemia is more likely when patients adopt this approach to diabetes self-management, which usually requires multiple daily insulin injections and maintenance of glycosylated hemoglobin levels within the normal range. Nonetheless, an increasing number of patients will probably embark on a program of tight glycemic con-

Wishy-washy terms in a single article

* I call these reviews, "trust me, I'm the expert" reviews.

† Oxman AD, Guyatt GH. The science of reviewing research. Authority, superstition, and science. Ann NY Acad Sciences 1993;703:125-33.

Evaluating Review Articles for Relevance and Validity

The goal of using this worksheet (jump to worksheet) is to quickly determine whether the review article is likely to present relevant and valid information. It allows determination, based on answers to the questions, whether the information is relevant to you, and whether the study design has sufficient rigor to apply the results to your patients.

The questions focus on the study design issues described above. The first 6 questions address study "musts" – they ask about issues of relevance or validity that must be present if the study results are to be applied to clinical practice. The answers to these questions must be yes regardless of the answers to the rest of the questions.

The goal of this worksheet is ***not*** to determine whether a study is "good" or "bad." Instead, we will use it to determine whether the results reported by the study are likely to occur if we use the same approach in our patients. As a result, the information is either useful to us, or not.

Step 1. How was relevant research identified?

Were the methods used to locate relevant studies comprehensive and clearly stated?

This question quickly separates summary reviews from synthesis reviews. The latter type of review will start by explaining how the authors assembled evidence for review.

"A Medline search was performed," is not an adequate explanation of a literature search. A Medline search will miss 30% - 50% of applicable controlled studies.

Instead, the methods should include:

1) A detailed explanation of the method used to search Medline, including search terms and strategies.
2) Searching of at least two databases. If the review involves a treatment, the Cochrane Central Register of Controlled Trials must be one of the searched databases. Other databases include
    a. The Health Technology Assessment Database
    b. EMBASE
    c. LILACS (Latin American and Caribbean Health Sciences Literature)
    d. DARE, The Database of Abstracts of Reviews of Effectiveness
    e. Web of Science
    f. Scopus
    g. HerbMed for botanicals
    h. Clinical trial registries. Most journals require that a study protocol be registered before the study is started, and a registry can identify studies that may have been completed but not published.
3) Unpublished and gray literature.

a. Results of studies may be published in meeting abstracts and, especially if a negative study, may not be published in a journal. The authors should look at the appropriate meeting abstracts for relevant research.
b. Not all research is published in journals but may be in government or other database.

4) Reference lists of identified articles. Since researchers will reference previous research in their own research descriptions, bibliographies are useful to find additional research.

Did they clearly outline study inclusion criteria that generalize to my practice?

The researchers should explain how they decided on which articles to include and exclude from their analysis. The included studies should include patients similar to patients in your care. For example, a meta-analysis of the effect of blood glucose control in intensive care surgical patients may not apply to patients on a medical ward.

Was the study selection independently performed by at least two investigators?

The literature search and article selection project should be performed by two researchers and their results compared.

Step 2. How valid was the identified research?

Garbage in = garbage out. There are several steps researchers must take to evaluate the research they have found:

a. Did the authors perform a validity assessment of the studies using appropriate criteria?

The assessment should be similar to those discussed for evaluating research regarding a therapy or diagnostic test.

b. Was the assessment independently performed by at least two investigators?

As mentioned above, the results should be compared and differences resolved, usually through discussion or by adjudication by a third researcher.

c. Were the included studies reasonably valid?

If not, how did the authors handle it? One option, which should be decided before the analysis is undertaken, is to include only studies of s certain quality level. A second approach is to separately analyze studies of high quality and low quality.

Step 3: Analyzing the data. Is it reasonable to combine these studies?

a. Were the included studies statistically homogenous?

Studies, conducted at different times, on different populations, with slight differences in design, will not produce the same results. The variability among studies' results is termed heterogeneity. Statistical heterogeneity occurs when the difference among study results is greater than chance alone.

Researchers will report evidence of heterogeneity (or lack thereof) in a couple ways:

1. Chi-squared test. Heterogeneity will produce a p-value <.05. Therefore, a higher p-value (e.g., > .05) is evidence of homogeneity
2. Degree of inconsistency (I^2):
    a. 0% to 40%: might not be important;
    b. 30% to 60%: may represent moderate heterogeneity;
    c. 50% to 75%: may represent substantial heterogeneity;
    d. 75% to 100%: considerable heterogeneity.
3. Below we will discuss a visual method of identifying heterogeneity.

b. If the results were heterogenous, is there a reasonable explanation?

Authors should try to identify a reason for the heterogeneity. It may be different study populations, study quality, etc.

Example

A meta-analysis compared administering an asthma drug by two different methods. The analysis found significant heterogeneity among the studies. The authors reasoned that the two methods might work differently in adults than in children. When they separated the results by age (children vs. adults), the heterogeneity was removed. The two administration methods were found to be equivalent in adults but not in children.

c. Were the populations, interventions, outcomes, and outcome measurements combined in a way that makes intuitive sense?

We often call this the "apples and oranges" issue. Meta-analysis only makes sense when combining results in a way that makes sense. A recent meta-analysis combined all "alternative medicine" approaches to treatment of a specific illness into a single analysis, which does not make intuitive sense.

d. Could publication bias have occurred?

Publication bias is the likelihood that negative results, i.e., studies that do not show a difference or benefit of a treatment, are less likely to be published.

Why?

1. Journal editors and journal reviewers are less interested in studies that "don't show anything."
2. A study is small, or smaller than previously published studies, and the results will not be interesting to readers.
3. Pharmaceutical companies suppress publication of research they paid for that isn't flattering (see TED talk, below, and a fascinating report of an example here).
4. Researchers do not submit research for publication because they know about #1.

So what?

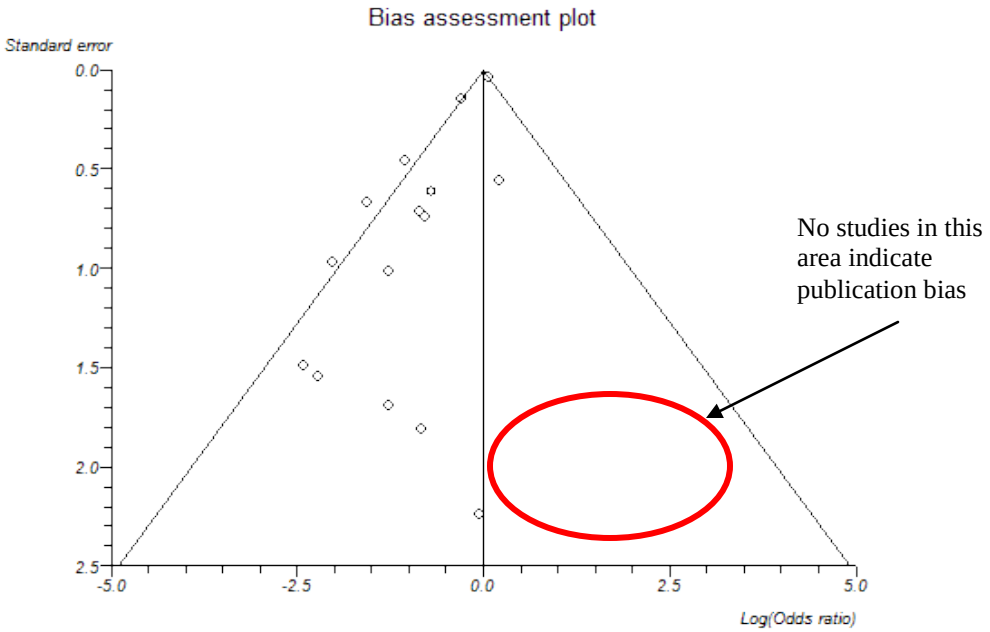
Since publication bias favors publication of research showing a benefit, a meta-analysis combining on published studies could inflate the real benefit of an intervention.

How to detect?

Researchers conducting meta-analysis can analyze the data to determine whether the risk of publication bias is high.

Statistically, different results from studies of the same topic should form a normal distribution (Gaussian curve) around the average calculated from those studies. That is, some results from individual studies should be below the mean, and some should be above. Also, the smaller the study, the greater the inherent variability of the data and the more likely the study is farther away from the mean.

A funnel plot compares the effect size in different studies with some measure of the variability of the data from each study. The example below compares the effect size as measured by odds ratio with the standard error (SE). Sample size is often used. Studies with small standard error cluster near the mean and studies with a larger SE are farther away. A "funnel" formed by the data that is balanced on both sides of the mean shows there was no publication bias. In the example that follows, the funnel is missing data at the bottom, left side, which is indicative of publication bias. Statistics can also be used to determine whether publication bias is likely.



Watch [TED Talk on the Effect of Publication Bias and Evidence Suppression] (13:29)

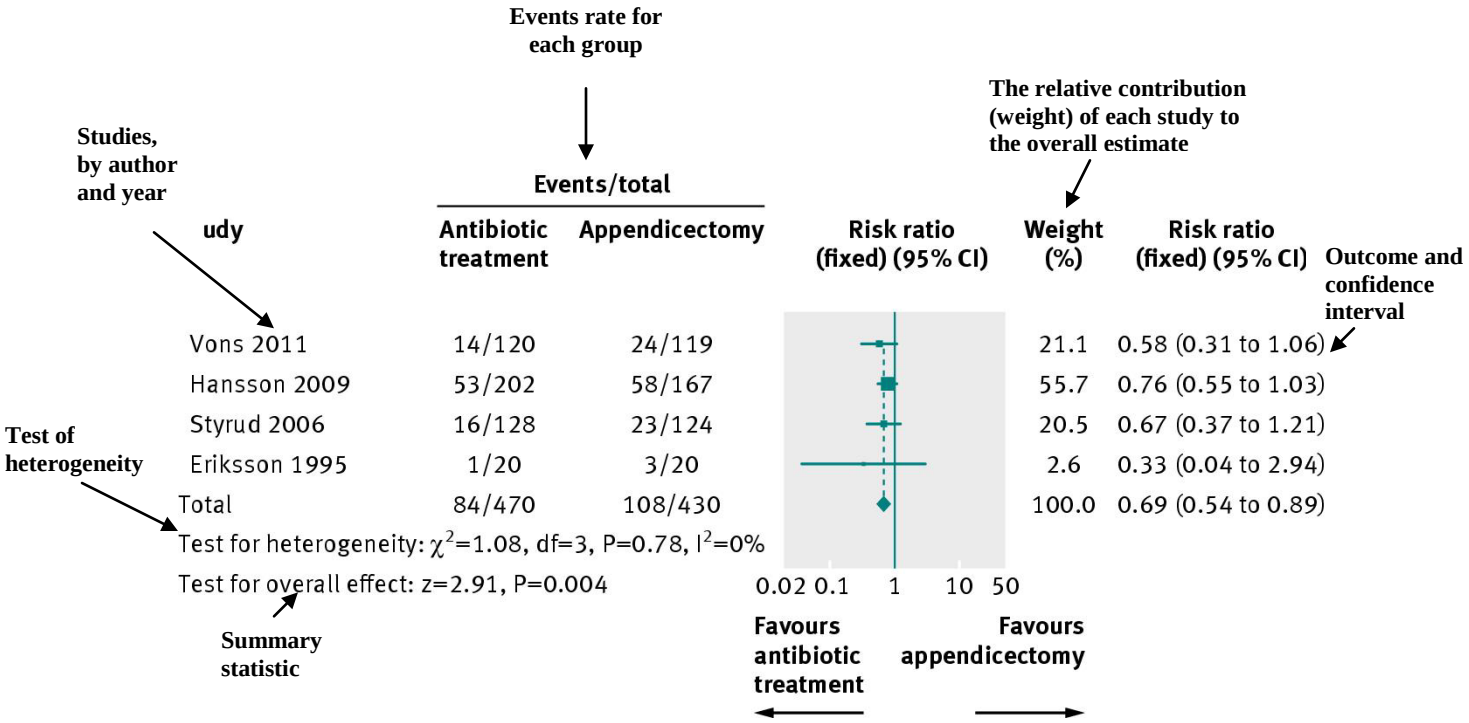
[Animation explaining publication bias] (3:02)

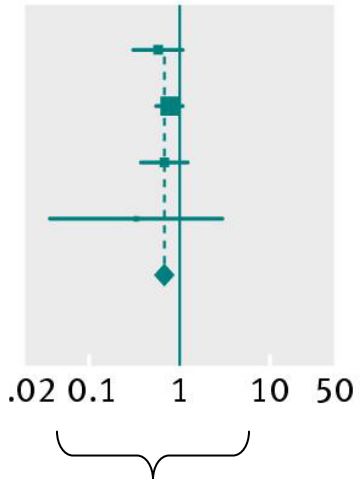
Interpreting the Box – Understanding the forest plot

The results of a meta-analysis are reported in a complex figure called a forest plot. It graphically illustrates the data from individual studies, their relative strength, and how the studies combine into a single result. The results of studies are reported in rows of numbers and graphically using a combination of boxes representing the result and horizontal lines representing the confidence intervals around the result. The resulting "forest of lines" is where the graph gets its name.

Read "Interpreting and Understanding Meta-Analysis Graphs"

A typical presentation from a meta-analysis in a forest plot:





Note that this is a log scale and differences get rapidly larger as the result moves from 1.

Each study's relative risk is presented as a box and horizontal line.

- The **box** represents the relative risk and its size conveys the relative weight of that study.
- The **horizontal bars** represent the confidence interval for that study.
- The **diamond** at the bottom is the result for the combined study results, with the horizontal points representing the confidence interval.

Interpretation:

- Study results with **horizontal bars** (confidence intervals) crossing the solid vertical line are not statistically significant.
- The **vertical dashed line** shows the relationship of combined result to the results for each study (heterogeneity). For this plot, there is little difference between the combined result and any of the individual studies.

Jump to instructions

Evaluating the Usefulness of Review Articles

Determine *Validity*

A. Finding the studies?

- Were the methods used to **locate** relevant studies comprehensive and clearly stated? ..Yes No
- Did they clearly **outline** study inclusion criteria that generalize to my practice?.....Yes No
- Was the study selection independently performed by at least **two** investigators? .....Yes No

E. **Validity**: Was the validity of the original studies appropriately assessed?

- Were the validity criteria **appropriate?** .....Yes No
- Was the assessment **independently** performed by at least two investigators?.....Yes No
- How were the **validity determinations** used?
    - If studies were *excluded,* were the criteria reasonable? .....Yes No
    - If all studies were *included,* did the authors perform a subanalysis based on study quality or sufficiently explain the influence? .....Yes No

F. **Analyzing** the Data: Is it reasonable to combine these studies?

- Were the studies **reasonably** valid? .....Yes No
- Were the included studies statistically **homogenous**? If not, did they provide an adequate explanation to account for the heterogeneity? .....Yes No
- Were the populations, interventions, outcomes, and outcome measurements **combined** in a way that makes intuitive sense? .....Yes No
- Could **publication bias** have occurred? .....Yes No

Study Guide
Session 5
Evaluating Clinical Practice Guidelines
Allen F. Shaughnessy, PharmD, MMedEd

The aims of this session are to:

1) Introduce you to the different types of clinical practice guidelines available to practicing physicians;
2) Present some issues that threaten the validity of recommendations in practice guidelines
3) Practice how to quickly and accurately evaluate clinical practice guidelines to determine their validity

Specific Objectives: By completing the initial reading and participating in class, students should be able to:

1) Quickly identify a guideline as being expert-based, evidence-based, or evidence-linked
2) Use the National Guideline Clearinghouse to find relevant clinical guidelines
3) Explain financial and intellectual conflicts of interest and how they can affect guideline recommendations
4) Use as set of questions to quickly evaluate a clinical practice guideline for threats to validity.

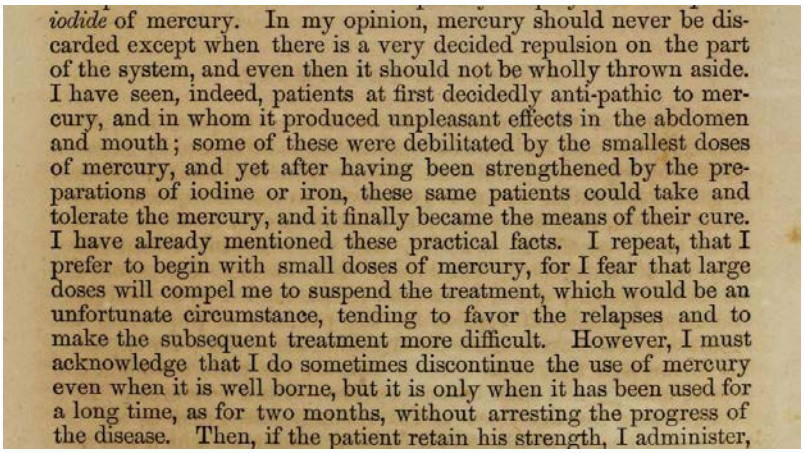
This study guide provides an outline of the concepts necessary to meet these objectives. It contains hyperlinks to short videos, web pages, or articles that explain the concepts in other ways or in greater detail. You can follow these hyperlinks if the explanations and examples I've given you are not sufficient to help you understand and to help you complete the readiness assessment test.

For more information, read (optional): Standards for Developing Trustworthy Clinical Practice Guidelines, Institute of Medicine or How NICE clinical guidelines are developed: an overview for stakeholders, the public, and the NHS.

What are clinical practice guidelines?

A **clinical practice guideline** aims to guide decisions by providing criteria regarding diagnosis, management, and treatment in specific areas of healthcare.

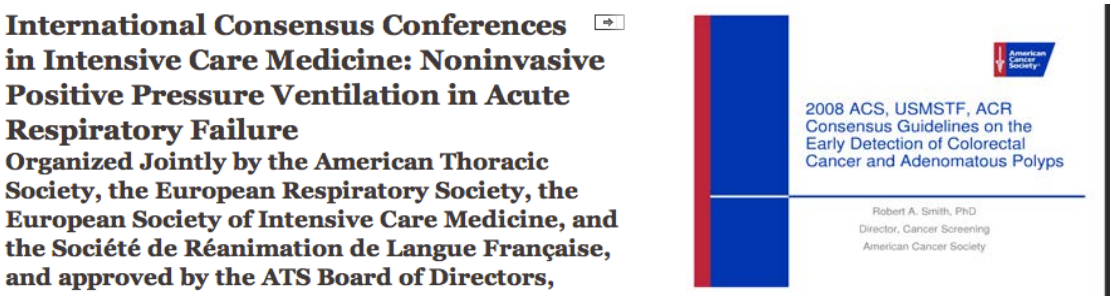
Guidelines have been produced throughout the history and have been based on tradition or authority. For example, a leading book on the treatment of sexually transmitted infections, written by a venerated venereal authority in 1859, states:



iodide of mercury. In my opinion, mercury should never be discarded except when there is a very decided repulsion on the part of the system, and even then it should not be wholly thrown aside. I have seen, indeed, patients at first decidedly anti-pathic to mercury, and in whom it produced unpleasant effects in the abdomen and mouth; some of these were debilitated by the smallest doses of mercury, and yet after having been strengthened by the preparations of iodine or iron, these same patients could take and tolerate the mercury, and it finally became the means of their cure. I have already mentioned these practical facts. I repeat, that I prefer to begin with small doses of mercury, for I fear that large doses will compel me to suspend the treatment, which would be an unfortunate circumstance, tending to favor the relapses and to make the subsequent treatment more difficult. However, I must acknowledge that I do sometimes discontinue the use of mercury even when it is well borne, but it is only when it has been used for a long time, as for two months, without arresting the progress of the disease. Then, if the patient retain his strength, I administer,

(Vidal, A-T. *A Treatise on venereal diseases.*)

There is a natural tendency to turn to experts for information and authority-based guidelines continue to flourish. They are often called "consensus guidelines." Frequently, this type of guidelines comes from professional groups or societies:



International Consensus Conferences in Intensive Care Medicine: Noninvasive Positive Pressure Ventilation in Acute Respiratory Failure
Organized Jointly by the American Thoracic Society, the European Respiratory Society, the European Society of Intensive Care Medicine, and the Société de Réanimation de Langue Française, and approved by the ATS Board of Directors,

The goal today is to have guidelines based on evidence, and most guidelines are now labeled as being evidence-based. *Evidence-based*, as we will discuss, does not necessarily mean "trustworthy."

Who produces clinical practice guidelines?

There are more than 3,700 guidelines from 39 countries in the Guidelines International Network database. These come from:

- Government agencies, e.g., The U.S. Preventive Services Task Force
- Medical associations, e.g., The American Thoracic Society
- Managed care organizations, i.e., insurers
- Other organizations, e.g., The Global Initiative for Chronic Obstructive Lung Disease
- Foundations, e.g., The National Osteoporosis Foundation
- Advocacy groups, e.g., The American Diabetes Association
- Commercial organizations
- *Ad hoc* groups, often funded by the pharmaceutical industry, e.g., treating pain due to peripheral neuropathy
- Local healthcare systems, e.g., Tufts Medical Center
- Individual practices

There are three general categories of guidelines:

1. Authority-based guidelines: BOGSATS

These are often called "consensus guidelines." Frequently, this type of guidelines comes from professional groups or societies. These are developed by bringing together a group of experts who decide how to write the guideline. Evidence is likely used by the group in some way, but there is no indication in the guideline regarding how they found, evaluated, and interpreted the evidence. As a result, the quality of the recommendations is low.

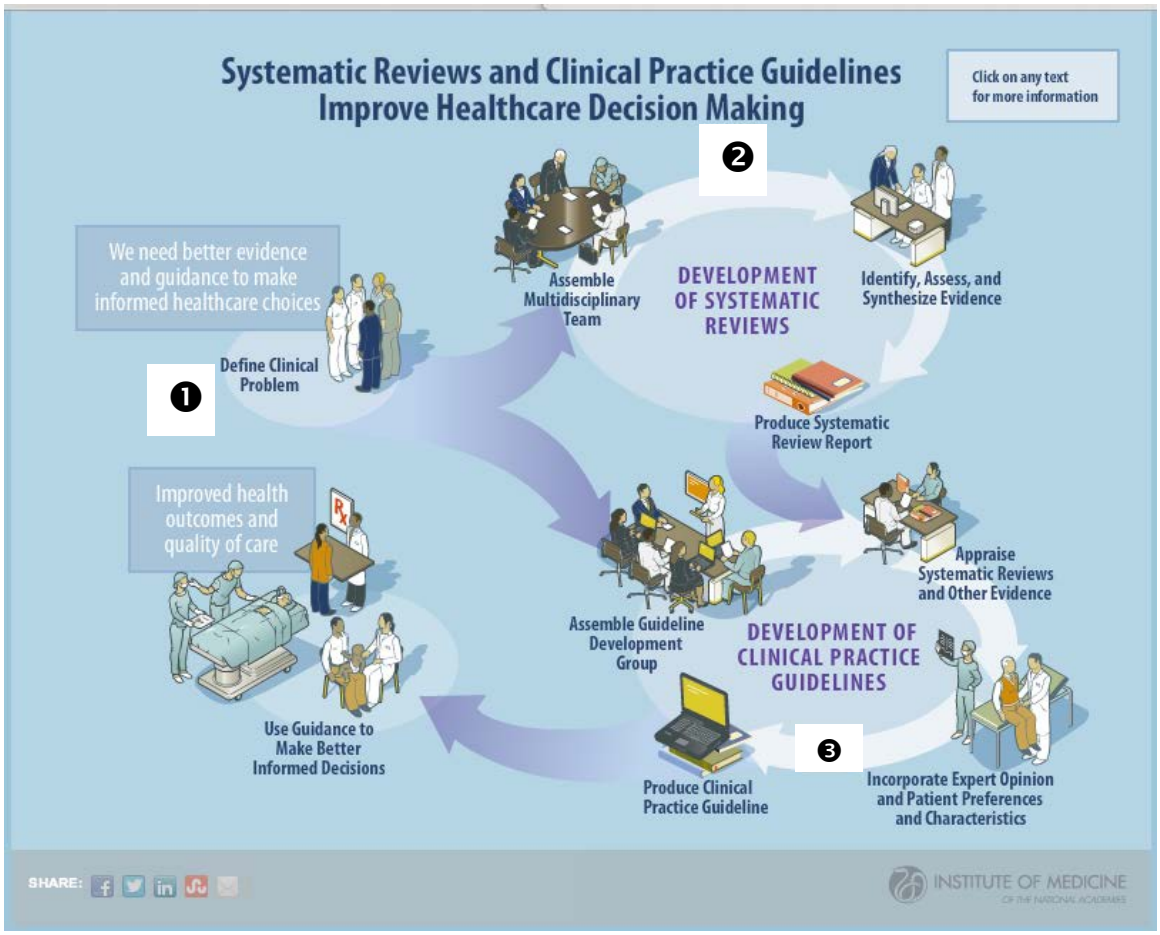
Less kindly, authority-based guidelines are called "**BOGSATS,**" which stands for a "bunch of old guys (or gals) sitting around a table."

"Consensus: A process by which a group agrees to something which no individual in the group believes to be appropriate."

-- Unknown

2. "Evidence-based" guidelines: Trust us, we have the evidence

The goal now is to have guidelines based on evidence, and most guidelines are now labeled as being "evidence-based." These guidelines start by identifying the clinical questions to be addressed by the guideline (❶, below). A group assembles all the pertinent evidence available to answer the questions into a systematic review (❷). A guideline development group evaluates the information, weighs the benefits and risks of different interventions, and develops a practice guideline (❸)

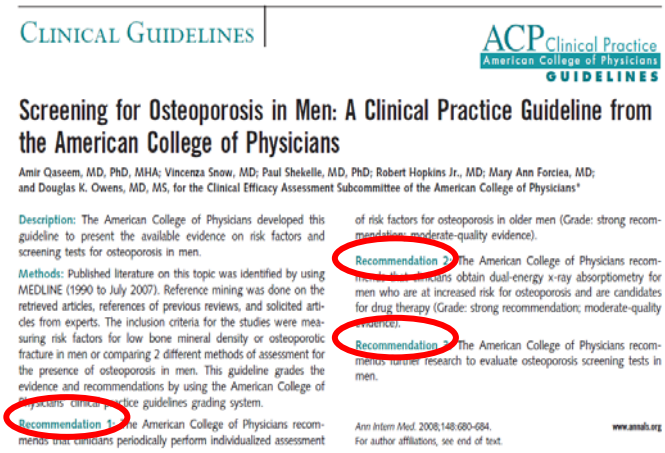


In contrast to evidence-*linked* guidelines, discussed below, evidence-*based* guidelines are limited by the lack of explicitness. They ask clinicians to be assured that they have used the evidence appropriately. Trust, though, is not part of EBM. Later in this section we will discuss how to identify guidelines written at this level.

3. Evidence-*linked* guidelines: Here is what we think and here is the evidence

Evidence-***linked*** guidelines can be identified by a methods section explaining how the evidence was determined, how it was graded, and how it was used to inform the decision-making behind the guidelines. Frequently this type of guideline has two documents: one document contains the recommendations that are linked to a second document that reports the evidence supporting these guidelines.

Recommendations



CLINICAL GUIDELINES

ACP Clinical Practice
American College of Physicians
GUIDELINES

Screening for Osteoporosis in Men: A Clinical Practice Guideline from the American College of Physicians

Amir Qaseem, MD, PhD, MHA; Vincenza Snow, MD; Paul Shekelle, MD, PhD; Robert Hopkins Jr., MD; Mary Ann Forciea, MD; and Douglas K. Owens, MD, MS, for the Clinical Efficacy Assessment Subcommittee of the American College of Physicians*

Description: The American College of Physicians developed this guideline to present the available evidence on risk factors and screening tests for osteoporosis in men.

Methods: Published literature on this topic was identified by using MEDLINE (1990 to July 2007). Reference mining was done on the retrieved articles, references of previous reviews, and solicited articles from experts. The inclusion criteria for the studies were measuring risk factors for low bone mineral density or osteoporotic fracture in men or comparing 2 different methods of assessment for the presence of osteoporosis in men. This guideline grades the evidence and recommendations by using the American College of Physicians clinical practice guidelines grading system.

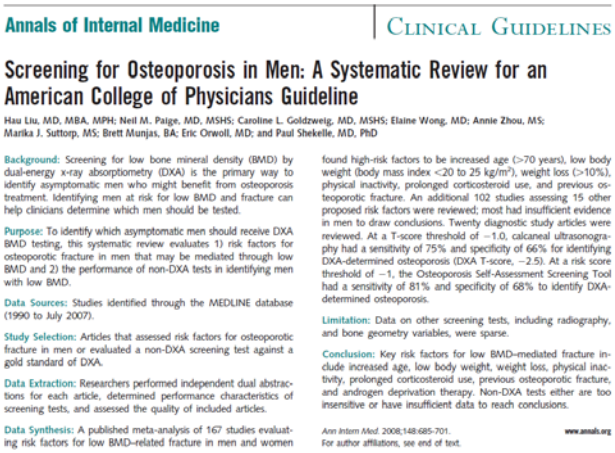
Recommendation 1: The American College of Physicians recommends that clinicians periodically perform individualized assessment of risk factors for osteoporosis in older men (Grade: strong recommendation; moderate-quality evidence).

Recommendation 2: The American College of Physicians recommends that clinicians obtain dual-energy x-ray absorptiometry for men who are at increased risk for osteoporosis and are candidates for drug therapy (Grade: strong recommendation; moderate-quality evidence).

Recommendation 3: The American College of Physicians recommends further research to evaluate osteoporosis screening tests in men.

Ann Intern Med. 2008;148:680-684. www.annals.org
For author affiliations, see end of text.

Evidence Report



Annals of Internal Medicine | CLINICAL GUIDELINES

Screening for Osteoporosis in Men: A Systematic Review for an American College of Physicians Guideline

Hau Liu, MD, MBA, MPH; Neil M. Paige, MD, MSHS; Caroline L. Goldzweig, MD, MSHS; Elaine Wong, MD; Annie Zhou, MS; Marika J. Suttorp, MS; Brett Munjas, BA; Eric Orwoll, MD; and Paul Shekelle, MD, PhD

Background: Screening for low bone mineral density (BMD) by dual-energy x-ray absorptiometry (DXA) is the primary way to identify asymptomatic men who might benefit from osteoporosis treatment. Identifying men at risk for low BMD and fracture can help clinicians determine which men should be tested.

Purpose: To identify which asymptomatic men should receive DXA BMD testing, this systematic review evaluates 1) risk factors for osteoporotic fracture in men that may be mediated through low BMD and 2) the performance of non-DXA tests in identifying men with low BMD.

Data Sources: Studies identified through the MEDLINE database (1990 to July 2007).

Study Selection: Articles that assessed risk factors for osteoporotic fracture in men or evaluated a non-DXA screening test against a gold standard of DXA.

Data Extraction: Researchers performed independent dual abstractions for each article, determined performance characteristics of screening tests, and assessed the quality of included articles.

Data Synthesis: A published meta-analysis of 167 studies evaluating risk factors for low BMD–related fracture in men and women found high-risk factors to be increased age (>70 years), low body weight (body mass index <20 to 25 kg/m^2), weight loss (>10%), physical inactivity, prolonged corticosteroid use, and previous osteoporotic fracture. An additional 102 studies assessing 15 other proposed risk factors were reviewed; most had insufficient evidence in men to draw conclusions. Twenty diagnostic study articles were reviewed. At a T-score threshold of −1.0, calcaneal ultrasonography had a sensitivity of 75% and specificity of 66% for identifying DXA-determined osteoporosis (DXA T-score, −2.5). At a risk score threshold of −1, the Osteoporosis Self-Assessment Screening Tool had a sensitivity of 81% and specificity of 68% to identify DXA-determined osteoporosis.

Limitation: Data on other screening tests, including radiography, and bone geometry variables, were sparse.

Conclusion: Key risk factors for low BMD–mediated fracture include increased age, low body weight, weight loss, physical inactivity, prolonged corticosteroid use, previous osteoporotic fracture, and androgen deprivation therapy. Non-DXA tests either are too insensitive or have insufficient data to reach conclusions.

Ann Intern Med. 2008;148:685-701. www.annals.org
For author affiliations, see end of text.

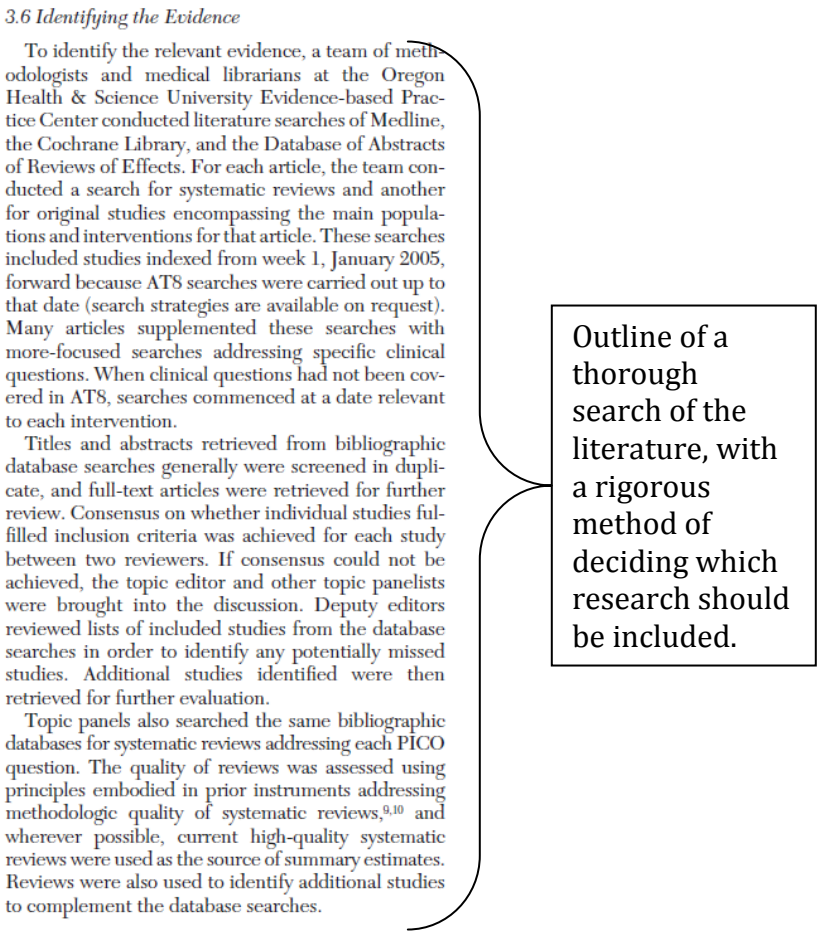
What are (potential) problems with guidelines?

Fundamentally, however, it is now nearly impossible for all stakeholders to be confident of [clinical practice guideline] quality.

-- Institute of Medicine, 2011, p. 191

Guidelines can differ markedly in their quality. A recent study showed that less than ½ of the 130 guidelines they evaluated met more than 50% of the requirements for good guideline development. Problems with guidelines include:

- **Lack of transparency**. There should be a clear path from the recommendations back to the evidence.
- **Guidelines that don't guide.** Sometimes, recommendations are written in a general form that makes it hard for clinicians to understand exactly what is being recommended. For example, a recent depression guideline suggested that, "it is not unreasonable to try exercise in certain patients."
- **Lack of systematic review of the literature.** As a result, guidelines sometimes suffer from selective citing of research that supports a certain position, ignoring research that doesn't support this position.



3.6 Identifying the Evidence

To identify the relevant evidence, a team of methodologists and medical librarians at the Oregon Health & Science University Evidence-based Practice Center conducted literature searches of Medline, the Cochrane Library, and the Database of Abstracts of Reviews of Effects. For each article, the team conducted a search for systematic reviews and another for original studies encompassing the main populations and interventions for that article. These searches included studies indexed from week 1, January 2005, forward because AT8 searches were carried out up to that date (search strategies are available on request). Many articles supplemented these searches with more-focused searches addressing specific clinical questions. When clinical questions had not been covered in AT8, searches commenced at a date relevant to each intervention.

Titles and abstracts retrieved from bibliographic database searches generally were screened in duplicate, and full-text articles were retrieved for further review. Consensus on whether individual studies fulfilled inclusion criteria was achieved for each study between two reviewers. If consensus could not be achieved, the topic editor and other topic panelists were brought into the discussion. Deputy editors reviewed lists of included studies from the database searches in order to identify any potentially missed studies. Additional studies identified were then retrieved for further evaluation.

Topic panels also searched the same bibliographic databases for systematic reviews addressing each PICO question. The quality of reviews was assessed using principles embodied in prior instruments addressing methodologic quality of systematic reviews,[9,10] and wherever possible, current high-quality systematic reviews were used as the source of summary estimates. Reviews were also used to identify additional studies to complement the database searches.

Outline of a thorough search of the literature, with a rigorous method of deciding which research should be included.

- **Conflicts of interest**. These can be financial or intellectual.
    - A financial conflict (duality) of interest can occur when a guideline developer receives research funding by a manufacturer of a product affected by the guideline, or if he or she is a paid speaker or consultant to a manufacturer or entity. A financial conflict of interest can also occur if a recommendation would result in financial benefit or harm to an organization or its members since a professional society has a primary responsible to promote its members' interests.[1] These financial arrangements may result in a conscious or unconscious bias.

Practice Guidelines: More Harm than Good? (watch only about the first three minutes)

    - Intellectual conflicts occur when one's research, profession, or experiences unduly influence one's ability to evaluate other types of evidence. It produces a sort of "tunnel vision." See déformation professionnelle.

KevinMD.com: *Conflicts of Interest Don't Always Involve Money*

How do these problems affect decision-making?

Clinicians are often confused when guidelines markedly differ from one another. For example, regarding the treatment of newborn infants with elevated bilirubin levels, the U.S. Preventive Services Task Force concludes:

> *"that the evidence is insufficient to recommend screening infants for hyperbilirubinemia to prevent chronic bilirubin encephalopathy."*

The American Academy of Pediatrics statement was published *the same month and year*, had the opposite recommendation:

> *"...We recommend universal predischarge bilirubin screening, which helps to assess the risk of subsequent severe hyperbilirubinemia. These recommendations represent a consensus of expert opinion based on the available evidence, and they are supported by several independent reviewers."*

[1] Quanstrum KH, Hayward RA. Lessons from the mammography wars. N Engl J Med 2010; 363:1076-1079.

Think about factors that might explain why these guidelines present such different guidance.

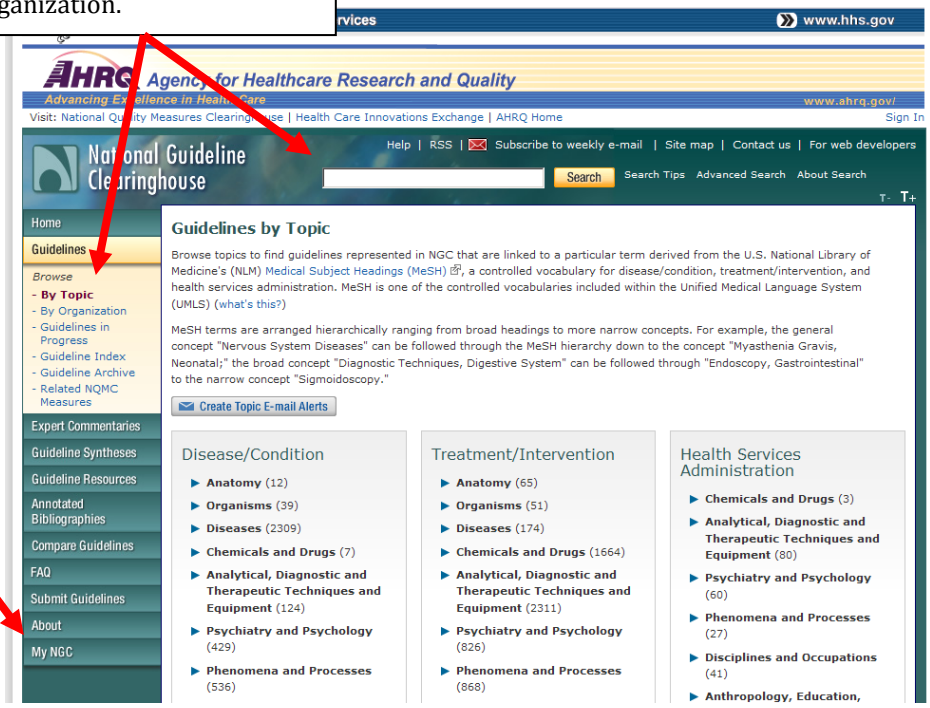
Finding Clinical Guidelines: The National Guideline Clearinghouse

http://ngc.gov/

The National Guideline Clearinghouse provides physicians and other health professionals, health care providers, health plans, integrated delivery systems, purchasers, and others an accessible mechanism for obtaining objective, detailed information on clinical practice guidelines and to further their dissemination, implementation, and use. The Clearinghouse actively searches for guidelines as well as accepts submissions from guideline development groups.

Guidelines can be searched by keyword, MeSH, or by topic or organization.

Users can create accounts and store guidelines or searches, as well as receive notification of updates.

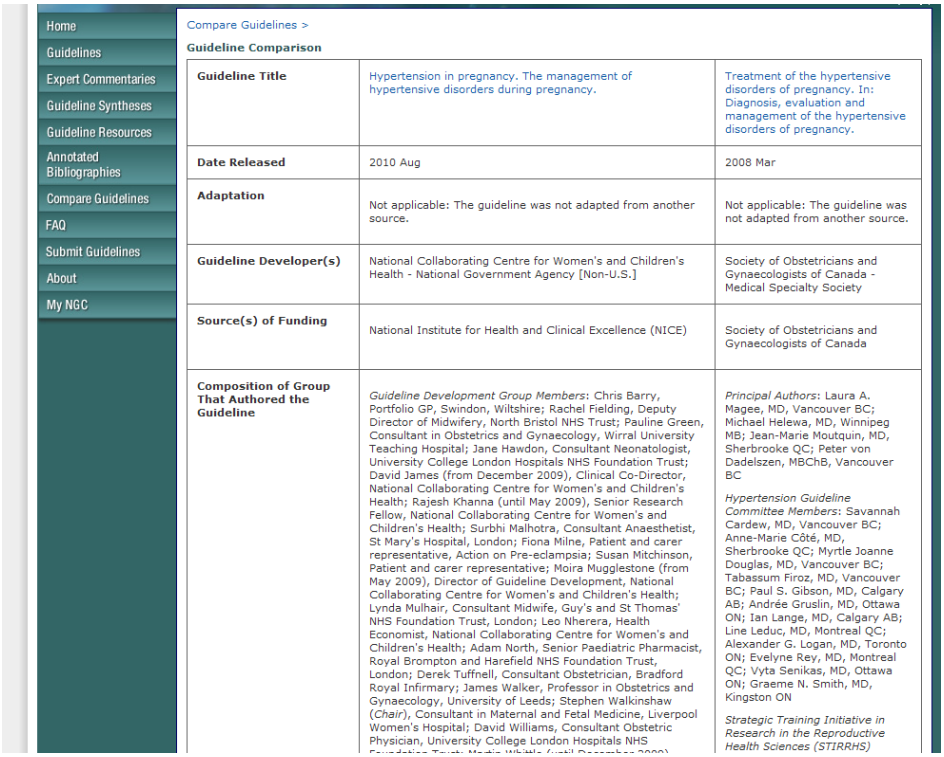


There is some quality control regarding which guidelines are included in the Clearinghouse. The NGC has specific criteria for including guidelines (these are in the process of being updated):

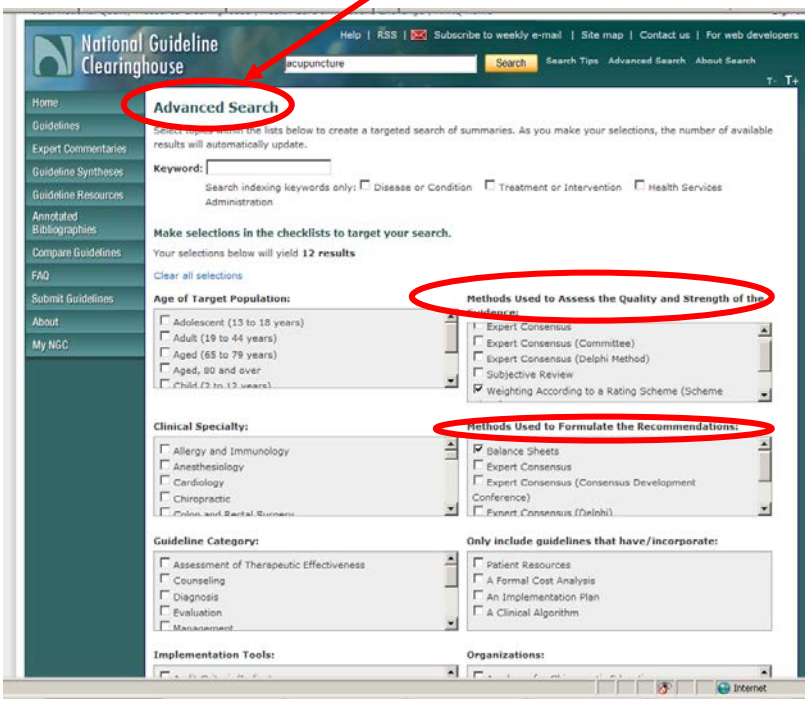
1) The clinical practice guideline contains systematically developed statements, strategies, or information for specific clinical circumstances;
2) The clinical practice guideline was produced under the auspices of medical specialty associations; relevant professional societies, public or private organizations, government agencies at the Federal, State, or local level; or health care organizations or plans.
3) There is evidence of a systematic literature search.

4) The full text of the guideline is available and has been developed, reviewed, or revised in the past 5 years.

Guidelines are summarized using a template, and guidelines can be **compared** to determine differences in methodology and recommendations.



Higher quality guidelines can be found by using the **Advanced Search** to limit results to guidelines with more methodologic rigor.



Evaluating Clinical Practice Guidelines

There are three characteristics of guidelines that must be evaluated:

1. The methodology used to identify and use the evidence.
2. The quality of the available evidence.
3. The presence of conflicts of interest.

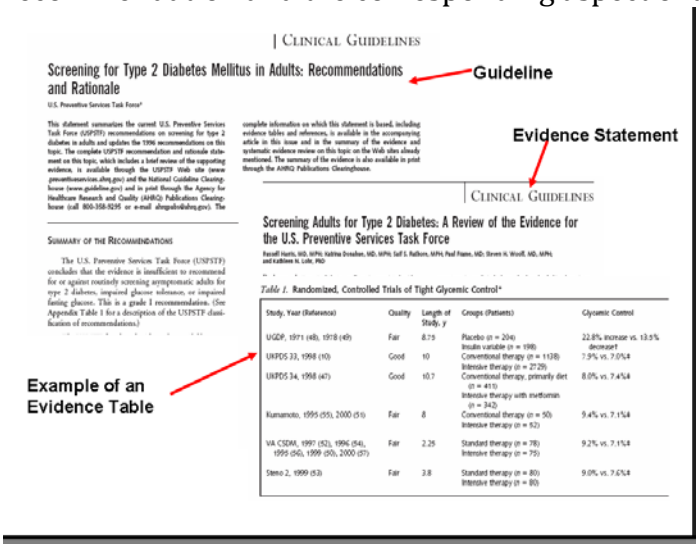
You can evaluate them by following the questions below. There is a worksheet that can be used to keep track of the answers. Often these questions can be answered by finding the appropriate section of the National Guideline Clearinghouse template.

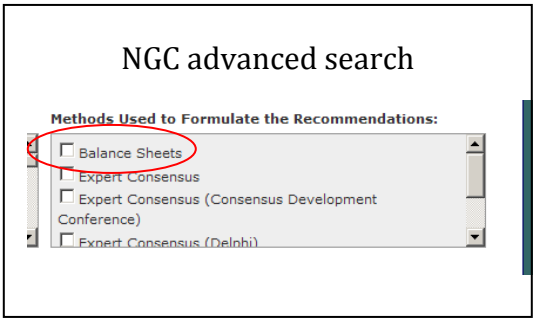
Evaluating the validity of the process: Identifying flaws in the methodology used to identify and use the evidence

The first step is to evaluate the process the guideline developers used. Evidence-based guidelines will describe an explicit process to finding, evaluating, and interpreting the evidence.

1. Evidence: Are the guidelines linked to a separate, systematic review of the evidence?

Recommendations in a guideline should be based on a systematic review performed by methodologists, not researchers in the field or the guideline development group. The guideline statements should be linked to this systematic review rather than requiring readers to trust the developers. This linkage can be quickly identified by finding *evidence tables*, *balance sheets*, or *indicators of the strength of recommendations*. Simply quoting selected evidence is not enough; there should be an explicit link between each recommendation and the corresponding aspect of the systematic review.





Balance sheet of benefits, risks and other aspects of decision-making

Strength of recommendation

Key Action Statement 5B

Clinicians may offer tympanostomy tubes for recurrent AOM (3 episodes in 6 months or 4 episodes in 1 year, with 1 episode in the preceding 6 months). (Evidence Quality: Grade B, Rec. Strength: Option)

Key Action Statement Profile: KAS 5B

Aggregate evidence quality	Grade B
Benefits	Decreased frequency of AOM. Ability to treat AOM with topical antibiotic therapy.
Risks, harms, cost	Risks of anesthesia or surgery. Cost. Scarring of TM, chronic perforation, cholesteatoma. Otorrhea.
Benefits-harms assessment	Equilibrium of benefit and harm.
Value judgments	None.
Intentional vagueness	Option based on limited evidence.
Role of patient preferences	Joint decision of parent and clinician.
Exclusions	Any contraindication to anesthesia and surgery.
Strength	**Option**

2. Chain of logic: Was an explicit, sensible, and transparent process used to weigh the risks and benefits associated with the recommendation?

Since all guidelines involve the application of values and judgments to the available evidence, this process should be described. The guideline should explain the target conditions, target populations, practice settings, and audience to which the recommendations apply. The outcomes should be specific – "clinically effective" is not a suitable outcome.

How the American College of Chest Physicians, sensibly, explicitly, and transparently selected outcomes and weighed risks and benefits.

3.2 Patient-Important and Surrogate Outcomes

The outcomes for each clinical question were chosen by the topic editors and their panel members and were generally consistent across articles. Outcomes were restricted to those of importance to patients.[4] Panels considered the burden of anticoagulation therapy as a patient-important outcome when its consideration could tip the balance of benefits and harms. If we found no data for an outcome considered at the outset as patient-important, we nevertheless included uncertainty about the effects of the intervention on that outcome when weighing its benefits and harms.

In the absence of data on patient-important outcomes, surrogates could contribute to the estimation of the effect of an intervention on the outcomes that are important. Examples of surrogate outcomes include asymptomatic venous thrombosis detected by venographic or ultrasound surveillance and the percentage of time that an international normalized ratio was in therapeutic range (used as a surrogate for bleeding and thrombosis in the assessment of the effectiveness of centralized anticoagulation services).

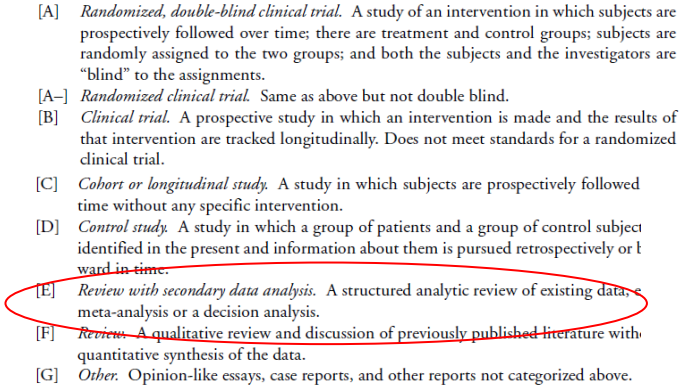
The issue of asymptomatic thrombosis detected by venographic or ultrasound surveillance presented particular challenges to the articles addressing VTE prevention in orthopedic and nonorthopedic surgery populations, an article addressing nonsurgical prophylaxis, and an article addressing stroke prevention. We were explicit in considering the trade-offs between VTE and bleeding events. An article by Guyatt et al[3] in this supplement addresses these issues in some detail.

Evaluating the Validity of the Supporting Research

3. Is the supporting research primarily randomized controlled trials, systematic reviews, or meta-analyses?

The next step is to review the evidence report to evaluate the quality of the research on which the guidelines are based. The guideline developers should describe their process for assigning levels of evidence. For example, some groups consider meta-analysis and systematic review to be lower quality evidence on the evidence hierarchy, a view not in line with current evidence-based medicine thinking. Guidelines based on low quality evidence should be clearly identifiable, either through levels of evidence or strength of recommendation ratings.

The American Psychiatric Association's evidence hierarchy, with meta-analysis near the bottom of the list.



[A] *Randomized, double-blind clinical trial.* A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are "blind" to the assignments.

[A–] *Randomized clinical trial.* Same as above but not double blind.

[B] *Clinical trial.* A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally. Does not meet standards for a randomized clinical trial.

[C] *Cohort or longitudinal study.* A study in which subjects are prospectively followed time without any specific intervention.

[D] *Control study.* A study in which a group of patients and a group of control subject identified in the present and information about them is pursued retrospectively or l ward in time.

[E] *Review with secondary data analysis.* A structured analytic review of existing data, e meta-analysis or a decision analysis.

[F] *Review.* A qualitative review and discussion of previously published literature with quantitative synthesis of the data.

[G] *Other.* Opinion-like essays, case reports, and other reports not categorized above.

Examples of good grading taxonomies:

United States Preventive Services Task Force Grades of Evidence

Strength of Recommendation Taxonomy (SORT)

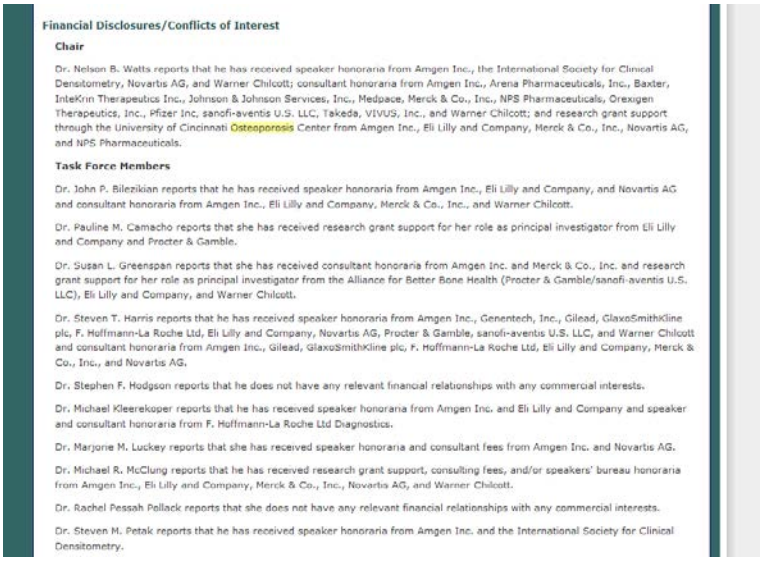
Evidence of Bias

As previously mentioned, intellectual bias, profession-based tunnel vision, and financial conflicts of interest can influence recommendations; especially those from professional advocacy groups. Conflicts of interest can have conscious or unconscious effects on the decision-making process. From physicians to U.S. Supreme Court Justices, most people with conflicts of interest do not recognize the effect on their judgments. Merely declaring or acknowledging conflicts does not mitigate their effects. Both social science and neuroscience literature demonstrate that transparency alone is an insufficient solution because bias is often implicit and unintentional. Moreover, disclosure may not only normalize conflicts of interest but may also worsen bias: "*disclosure can actually lead doctors to give biased advice, either through strategic exaggeration (whereby more biased advice is provided to counteract anticipated discounting), or "moral licensing" such that advice is legitimized because advisees "have been warned" (that is, caveat emptor or "buyer beware").*"

4. Financial conflict of interest: Are most of the guideline developers free of declared financial conflicts of interest, especially the committee's chair?

Guidelines should contain a statement explaining the financial conflicts of interest of the guideline developers. The Institute of Medicine suggests that a minority of the development group should have conflicts and that the guideline chair should not have any financial ties.

Example of a guideline development group with most members declaring conflicts of interest, including the chair.



Financial Disclosures/Conflicts of Interest

Chair

Dr. Nelson B. Watts reports that he has received speaker honoraria from Amgen Inc., the International Society for Clinical Densitometry, Novartis AG, and Warner Chilcott; consultant honoraria from Amgen Inc., Arena Pharmaceuticals, Inc., Baxter, InteKrin Therapeutics Inc., Johnson & Johnson Services, Inc., Medpace, Merck & Co., Inc., NPS Pharmaceuticals, Orexigen Therapeutics, Inc., Pfizer Inc, sanofi-aventis U.S. LLC, Takeda, VIVUS, Inc., and Warner Chilcott; and research grant support through the University of Cincinnati Osteoporosis Center from Amgen Inc., Eli Lilly and Company, Merck & Co., Inc., Novartis AG, and NPS Pharmaceuticals.

Task Force Members

Dr. John P. Bilezikian reports that he has received speaker honoraria from Amgen Inc., Eli Lilly and Company, and Novartis AG and consultant honoraria from Amgen Inc., Eli Lilly and Company, Merck & Co., Inc., and Warner Chilcott.

Dr. Pauline M. Camacho reports that she has received research grant support for her role as principal investigator from Eli Lilly and Company and Procter & Gamble.

Dr. Susan L. Greenspan reports that she has received consultant honoraria from Amgen Inc. and Merck & Co., Inc. and research grant support for her role as principal investigator from the Alliance for Better Bone Health (Procter & Gamble/sanofi-aventis U.S. LLC), Eli Lilly and Company, and Warner Chilcott.

Dr. Steven T. Harris reports that he has received speaker honoraria from Amgen Inc., Genentech, Inc., Gilead, GlaxoSmithKline plc, F. Hoffmann-La Roche Ltd, Eli Lilly and Company, Novartis AG, Procter & Gamble, sanofi-aventis U.S. LLC, and Warner Chilcott and consultant honoraria from Amgen Inc., Gilead, GlaxoSmithKline plc, F. Hoffmann-La Roche Ltd, Eli Lilly and Company, Merck & Co., Inc., and Novartis AG.

Dr. Stephen F. Hodgson reports that he does not have any relevant financial relationships with any commercial interests.

Dr. Michael Kleerekoper reports that he has received speaker honoraria from Amgen Inc. and Eli Lilly and Company and speaker and consultant honoraria from F. Hoffmann-La Roche Ltd Diagnostics.

Dr. Marjorie M. Luckey reports that she has received speaker honoraria and consultant fees from Amgen Inc. and Novartis AG.

Dr. Michael R. McClung reports that he has received research grant support, consulting fees, and/or speakers' bureau honoraria from Amgen Inc., Eli Lilly and Company, Merck & Co., Inc., Novartis AG, and Warner Chilcott.

Dr. Rachel Pessah Pollack reports that she does not have any relevant financial relationships with any commercial interests.

Dr. Steven M. Petak reports that he has received speaker honoraria from Amgen Inc. and the International Society for Clinical Densitometry.

http://ngc.gov/content.aspx?id=34968&search=osteoporosis

5. Intellectual conflict of interest: Are the developers from a range of specialties, and include patients and other stakeholders?

Guidelines, especially from specialty societies are at risk of an intellectual conflict of interest when all group members are from the same profession and are members of the professional society promoting and developing the guidelines.
See: **http://tinyurl.com/d5gtznv.**

Even if the guidelines are not self-serving, there is a tunnel vision that occurs when all guideline developers are looking at evidence through the same prism of experience. Guideline development groups should have representatives from different specialties and, where possible, patients or consumer advocacy groups. Ideally, a methodologist should be part of the group as well.

Active researchers, while a source of cutting edge information, also risk being unduly influenced by their own research findings to the exclusion of other evidence. Evaluating for this bias can be accomplished by checking the author affiliations to determine whether they are active researchers, as well as by searching the guideline citations for publications by guideline development group members.

***6. Professional Conflict of Interest*:** Were the guidelines approved or modified by a board, executive committee, or consensus committee of a professional society before their release?

Professional societies may not be able to provide unbiased guidance. As mentioned in a recent editorial,

> ". . . Although it is true that individual medical providers care deeply about their patients, the guild of health care professionals — including their specialty societies — has a primary responsibility to promote its members' interests. . . But it is a fool's dream to expect the guild of any service industry to harness its self-interest and to act according to beneficence alone — to compete on true value when the opportunity to inflate perceived value is readily available."

Recommendations are suspect if they are written by a specialty group and the recommendations would benefit that specialty group's members. To quote an old adage, "Never ask a barber if you need a haircut."

Back to instructions

A Worksheet for Evaluating Practice Guidelines

Determine ***Relevance***

Is this guideline worth considering? If the answer to any of these questions is No, it may be better to read other articles first.

A. Are the recommendations based on research on outcomes that patients would care about? (Be careful to avoid results that require extrapolation to an outcome that truly matters to patients)

Yes (go on) No **(stop)**

B. Does the guideline address problems common to your practice and suggest feasible interventions?

Yes (go on) No **(stop)**

C. Will this information, if true, require you to *change* your current practice?

Yes (go on) No **(move on to the next guideline)**

Back to instructions

Determine ***Validity*:**

D. Validity of the process:

1. *Evidence*: Are the guidelines linked to a separate, systematic review of the evidence? Yes No

2. *Chain of logic*: Was an explicit, sensible and transparent process used to weigh the risks and benefits associated with the recommendation? Yes No

E. Validity of the supporting research

3. *Validity:* Is the supporting research primarily randomized controlled trials, systematic reviews, or meta-analyses? Yes No

F. Evidence of Bias

4. *Financial conflict of interest*: Are most of the guideline developers free of declared financial conflicts of interest, especially the committee chair? Yes No

5. *Intellectual conflict of interest*: Are the developers from a range of specialties, and include patients and other stakeholders? Yes No

6. *Professional Conflict of Interest*: Were the guidelines approved or modified by a board, executive committee, or consensus committee of a professional society before their release? Yes No

Back to instructions