EVIDENCE-BASED MEDICINE

"I use not only all the brains I have, but all I can borrow."

Woodrow Wilson

Introduction

Two 'pillars' should sustain primary care medicine, evidence-based knowledge and authentic human interaction. Truth in these two areas is grounded in different paradigms, or ways of understanding the world. Evidence-based medicine (EBM) focuses on objective observations, explanations and causative algorithms. Authentic human interaction is based on subjective meaning, respect, and value sharing. Elements from the latter paradigm are often difficult or impossible to frame within the former.

For evidence-based knowledge, it is not enough to know the medical facts of today. Medical knowledge is changing too rapidly. You must also develop a strategy for keeping up with these advances and assessing the 'truth' of your current understanding for a given disease, its diagnosis and treatment. Evidence comes from quantitative research methods such as randomized trials and observational studies. In this manuscript, we emphasize a five-step approach to evaluating this evidence.

If you are going to continue the habit of 'keeping up', it will only be because you have created and practiced a useful method. We cannot overemphasize the importance of developing these skills now.

Learning Goals

We hope that reading the literature to answer your clinical questions will become second nature to you. To meet this goal, we hope you:

- Develop a personal approach to scanning the medical literature.
- Understand the types of research studies available, how their results are reported, and which types of statistical treatments are appropriate for which types of data.
- Understand what factors may be responsible for a falsely positive or falsely negative result in a clinical trial, and be able to evaluate these factors.
- Can take a valid result from a clinical trial and describe what that result means for your individual patient.

Qualitative Research

The function of qualitative research is to create hypotheses, unlike the more familiar model of experimental research, which is designed to test hypotheses. Qualitative research is often used in areas such as anthropology and sociology because the systems being studied are too complex, interdependent and historically bound to be encompassed by an easily articulated hypothesis. Ask yourself what hypothesis you would test to find out why you decided on the specialty you have chosen? At a first pass, it would probably include a complex match between types of patients and problems, and your cognitive style and values. What about your experience during training? Your role models? The amount of autonomy you were granted during training? The current political pressures on the medical system? These have all been the subjects of interesting qualitative research in medical education. The results of qualitative research are basically reported, in greater or lesser detail, as a story. Its validity is only for the specific situation, setting and time studied, but it may lead to deeper more generalizable hypotheses that can be tested.

Quantitative Research

Quantitative research and EBM test hypotheses. They answer who, what, when, where, why, how, and how much questions in medicine. There are two important assumptions that are frequently overlooked in the traditional EBM paradigm. First, EBM relies on specifiable variables, rigorous study design, and cause-effect inferences. This means that some relevant differences cannot be easily studied.[1] For instance, how would you randomize a patient's willpower? The second assumption is that EBM views each article as an independent search for truth. This requires critical review of all potentially pertinent articles. Both of these assumptions affect the value of traditional EBM.

CAUSE-EFFECT INFERENCES

How do we prove that something (say A) is the cause of something else (say B)? In general, we need to satisfy three criteria:

- correlation (A and B occur together)
- sequence (A always occurs before B)
- isolation (A is the only variable that could have caused B)

Observational studies address the first two criteria but do not establish causation Observational studies may be the only way to study some problems because of ethical considerations or cost (e.g., smoking as a cause of lung cancer will never be randomized). Observational studies tend to be cheaper and are often used to identify the promising correlations that should be studied further. For example, the Framingham trial established the hypothesis that hypertension is a risk factor for CAD. It was followed by several randomized treatment trials, which tested and proved that hypothesis.

Randomization is the critical (and costly) ingredient that makes something an experimental study. Experimental studies are the only way to isolate the effect from other potential causes and fully establish causation. To state it another way, randomization should evenly balance all other potential causes (whether we know about them or not) except the one being studied. In the early hypertension treatment trials mentioned above, we didn't know that we might need to measure C reactive protein or chlamydia titers. Now, we assume they were balanced by the randomization because the other known risk factors (smoking, lipids, etc.) were adequately balanced.

This need to define, isolate and control EBM experiments has a potential down side. In general, the more homogeneous the population and the more scientifically sound the study, the more restricted is the conclusion for the general population, and vice versa.

ARTICLE 'TRIAGE' PHILOSOPHY

As mentioned above, traditional EBM views each article as an independent search for truth. This requires full evaluation of any potentially important article. We recommend a modified approach.

We assume that you are busy; that you have an established practice for the common and important clinical questions; and that you would be open to changing your practice if a study provided a good reason to do so. This difference in philosophy is at the heart of our approach's efficiency. Using a legal metaphor, traditional EBM requires each article's conclusion to be 'true beyond a reasonable doubt', a strict criminal law standard. Our approach weighs each article's conclusion with existing knowledge and practice to come to "a preponderance of the evidence", a flexible civil law standard. It is founded in a Baysian (rather than p values) approach to truth and validity.[2, 3]

The five-step method

The five-step method shown in figure 1 is an efficient way to eliminate articles that are not likely to change or support current practice.

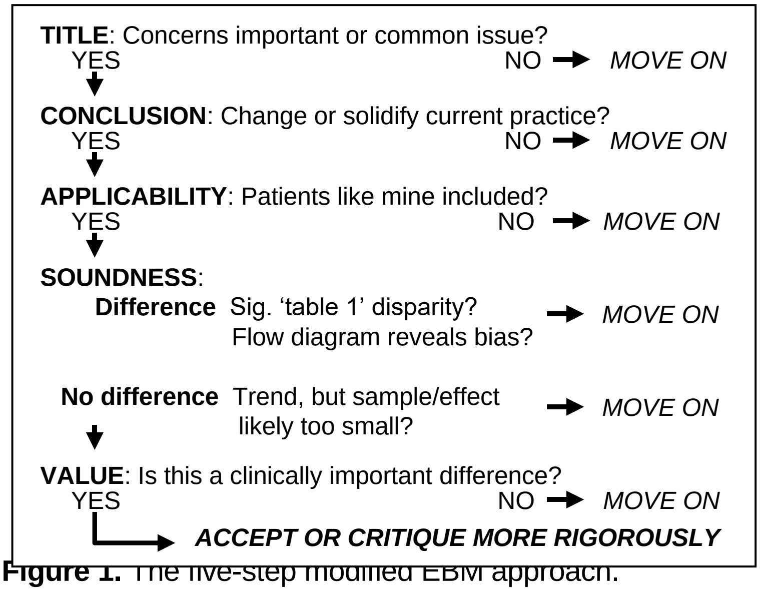

TITLE: Concerns important or common issue?
YES
NO
MOVE ON
CONCLUSION: Change or solidify current practice?
YES
NO
MOVE ON
APPLICABILITY: Patients like mine included?
YES
NO
MOVE ON
SOUNDNESS:
Difference
Sig. 'table 1' disparity?
Flow diagram reveals bias?
MOVE ON
No difference
Trend, but sample/effect likely too small?
MOVE ON
VALUE: Is this a clinically important difference?
YES
NO
MOVE ON
ACCEPT OR CRITIQUE MORE RIGOROUSLY

Figure 1. The five-step modified EBM approach.

Types of Quantitative studies

OBSERVATIONAL STUDIES

Observational studies, as their name implies, are descriptive studies based on passive observation of characteristics and diseases. In general, they assess risks. They can show promising correlations, but generally do not reach causal conclusions. The common observational study classes are designed to answer one of three questions.

- What is happening? (cross-sectional)
- What did happen? (case-control)
- What will happen? (cohort)

Cross-sectional studies are done at a single point in time. The assignment to groups of interest and the assessment of the characteristics of interest occur simultaneously. These studies are generally used to characterize a study group, or to evaluate a diagnostic procedure. For example, we might design a questionnaire to be filled out on the oncology ward in patients with colon carcinoma, and in matched controls, to find out if they had prior flexible sigmoidoscopy. This would tell us how many more times likely it was that those with carcinoma had missed being screened.

In a case-control design, we identify the outcomes (had colon cancer or control in the case above) and then look retrospectively at the risks or exposures. As opposed to cross-sectional designs, we may interact with the patients several more times about occurrences prior to the outcome. For instance, we might find that a flexible sigmoidoscopy performed by a generalist or a surgeon is not as reliable as one performed by a gastroenterologist. We might then wish to re-contact all positive responders from the above study to find out who performed the procedure. We may also wish to review the charts of some percentage of those with a negative response to assure that they did not have a sigmoidoscopy. Case-control studies are particularly useful for documenting rare diseases, examining the effect of low-frequency treatments, or searching for causal implications (to be further tested). Again, we may only infer how many more times likely it was that those with the disease (cancer) had the characteristic (lack of screening).

Cohort studies are prospective. You identify a likely candidate "cause" for some "effect" you are interested in, and stratify the exposed and unexposed groups. You then follow them over time. In the above example, we might look at all those who either receive or do not receive a sigmoidoscopy, and wait to see who develops cancer. How is this different than a randomized controlled trial? The most important difference is that first word, *randomized*. By allowing the medical care system to decide who gets screened or not, we may be naturally selecting a group with a different outcome. For instance, if we feel that when it is likely that someone has cancer, we should skip sigmoidoscopy and go directly to colonoscopy, then patients in the "no sigmoidoscopy" group may be generally healthier. This is not because screening made a difference, but because a selection bias led to a difference between the groups. Randomization does two things. It decreases the potential for our own biases to affect outcome, and it equally stratifies other unrecognized factors which may be important (family history of colon cancer in our above example). Cohort studies are good for searching for causal factors, and can tell you how many times more likely it is that patients with a given characteristic will become diseased.

EXPERIMENTAL STUDIES AND META-ANALYSES

Experimental studies are those in which the investigator manipulates an intervention by assigning or withholding it to randomized groups, and then compares the outcomes. The epitome is the randomized, placebo-controlled, blinded clinical trial. Randomized, placebo-controlled trials involve selection of a group of eligible patients based upon predetermined criteria, random allocation to either the study or a "sham" intervention, and control of extraneous variables that might confound results. These studies are useful for comparing effects of various therapies on health outcomes, and for inferring cause. As we have stated earlier, they have high internal validity.

Meta-analyses are large groupings of experimental studies looking at similar questions. They are largely done to improve statistical power. Thus, we might have 20 small studies of 50 patients each to examine the sigmoidoscopy screening question. If screening lowered the incidence of cancer by 30%, we would likely not detect this finding within any one of these studies. However, lumping them all together as one study with 1000 patients may show a 30% reduction. The two most common problems with meta-analyses are that different studies have different baseline characteristics (you may be lumping apples with oranges), and that negative trials are somewhat less likely to be reported, and so the more you lump, the more you favor a positive result (publication bias).

Further EBM Analysis

Now that we have discarded articles that are not likely to change or support our current practice, we are left with studies that are important for our patients. It is useful to evaluate them in a systematic fashion. We favor an approach used by many journals, which we call "ODSPIRC"ing. This is not a simple exercise in ‘filling in the blanks’. Rather, it should be a value judgment in each category about the adequacy and appropriateness of this element for the clinical question.

ODSPIRC is:

O bjective Hypothesis being tested (stated in terms of the intervention and the primary endpoints of the study).

D esign See levels of evidence on page II-3 of this syllabus.

S etting What nation, hospital/clinic type, and expertise are involved.

P atients Inclusion and exclusion criteria, particularly with respect to other prognostic factors.

I ntervention What type? What dose? For what length of time?

R esults How do they apply to an individual patient? Do they make sense given the hypothesis? What is the number needed to treat or harm (see below)?

C onclusions Clinically significant and achievable in my setting?

We now have some basis for our thumbs up or thumbs down decision on any suggested change in our practice.

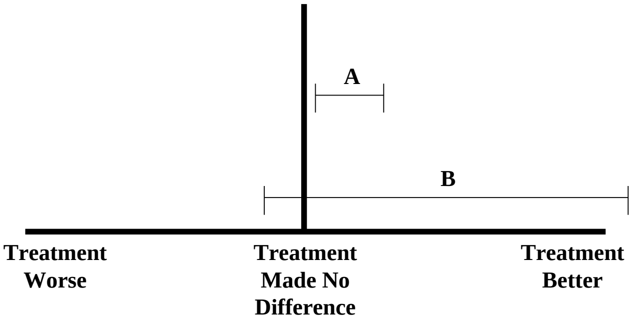
MORE ABOUT THE MATH

At this stage we have decided that an article will be very useful if its results are accurate. Now we need some skills in deciding 'this is the truth'. Many of the articles of interest will be clinical trials (a type of experimental study). In these clinical trials we use a statistical tool called the *null hypothesis*. Simply stated, we make the assumption that there *will not be* a difference between two populations (treated and not treated for example) caused by our intervention. If a difference is found, it is assumed to be due to our intervention.

What we really want to know is:

- Are the two groups truly different?
- What is the magnitude of the difference?
- What is the precision of our estimate of the two groups?

One way of looking at this is to draw a graph showing whether the intervention had a positive, a negative or no (null) effect. If the data includes the centerline (the null hypothesis) then it does not achieve statistical significance. The limits right and left are the confidence intervals, usually 95%. These limits are difficult to explain exactly. However, if several studies are done, 95% of the 95% confidence intervals will include the "true" value for the population as a whole.



The P-value expresses whether a study result achieves statistical significance or not. It is defined as the possibility that chance alone produced these results (usually considered significant if < 0.05). P-values alone however may be misleading.[2, 3, 4] In the example pictured, treatment A does not include the null hypothesis, and is therefore statistically significant. It, however, may not be a very significant effect clinically (doesn't change things much from null). Treatment B does include the null and is therefore not statistically significant. It may, however, have a greater chance of having a bigger clinical difference. You can see that knowing the confidence intervals adds a lot to your understanding of both positive and negative studies.

Another example:

	Placebo	Treatment
Success	7	14
No success	13	6
	20	20

Here p = 0.06 (NOT significant). However, the placebo has only 7/20 = 35% success, while the active treatment has 14/20 = 70% success. The absolute difference with treatment is 35% and may be clinically significant (incidentally, the 95% confidence intervals above are from -1% to 71%).

OBSERVATIONAL STUDIES

Observational studies do not use the null hypothesis but, rather, deal with risk. When measuring *the degree to which two variables are associated*, cross-sectional and case-control designs should report odds ratios. Odds ratios are retrospective. They answer the question: How many times more likely is it that a diseased individual has the characteristic? Observational studies should report relative risk. Relative risk is prospective. It answers the question: How many times more likely are people with the characteristic to become diseased?

NUMBER NEEDED TO TREAT (OR HARM)

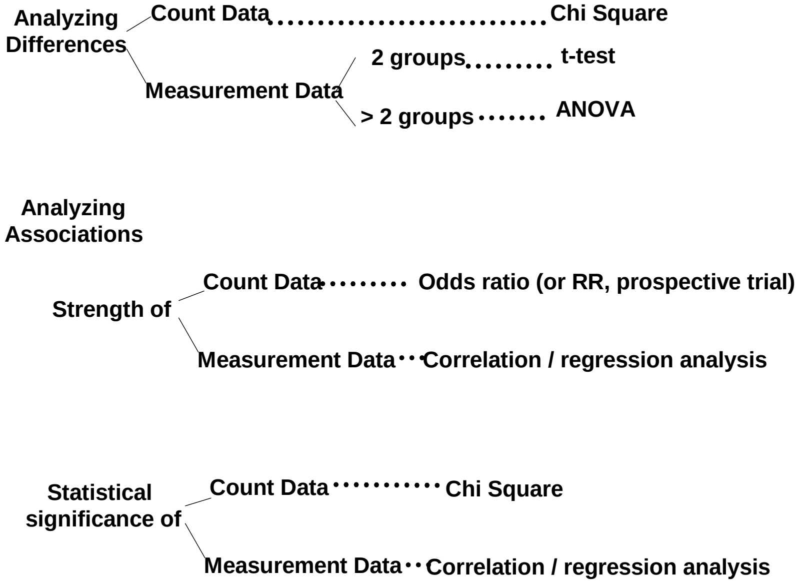
Given a study that appears to have a high degree of internal validity and an acceptable degree of external validity, how useful will it be for an individual clinic patient? We now introduce the concept of number needed to treat. The number needed to treat is calculated as the absolute risk reduction (in %) divided into 100. If we have a study of 100 patients where 10 of the control group had a certain outcome and 5 of the intervention group had it, then their relative risk reduction (RRR) would be [(10 – 5) ÷ 10] x 100 = 50%, and their absolute risk reduction (ARR) would be (10

– 5) = 5%. Dividing this ARR into 100 yields a number needed to treat of 20.

As an example, let's examine the CAPRIE trial (Lancet 1996:348:1329). This was a study of 19,185 patients looking at aspirin vs clopidogrel for prevention of combined CVA, MI and vascular death (primary outcome measure). After an average of 1.9 years, the clopidogrel arm had a statistically significant lower event rate (5.32%) than the aspirin arm (5.83%, p = .043). This represented a relative risk reduction of 8.7%. Does it seem like a good thing in someone who has failed aspirin?

While the relative risk reduction was 8.7%, the absolute risk reduction (ARR = .0583-.0532) was only .0051 or about ½%. Thus, the NNT is 196. This means you would have to treat 196 patients for 1.9 years to prevent one endpoint (CVA, MI, or death). At our institution, you would have to spend $ 225,000 per year to prevent one endpoint!

SELECTING THE CORRECT ANALYTICAL METHOD
Here are some rules of thumb for deciding if the correct analytical method was used in an article. Count data is nominal (e.g., male/female), or ordinal (e.g., a '1-5' number circled on a questionnaire, where there is direction but the differences aren't always equal—the difference between 1 and 2 on the questionnaire may not equate to the difference between 4 and 5). Measurement data is usually interval (e.g., temperature, weight—there is direction and a change of one unit anywhere on the scale is equal).



If all of this seems a little too mathematical, there are two excellent references that use primarily pictorial and story explanations of these concepts. The "statistics 101" version, suitable for all physicians, is *PDQ Statistics*.[5] The "statistics 201" version, useful if you are planning on research and/or a fellowship, is the *Primer of Biostatisitics*.[6]

Bibliography

1. Tonelli MR. The Philosophical Limits of Evidence-based Medicine. Acad Med 1998;73:1234-1240.

2. Goodman SN. Toward Evidence-Based Medical Statistics. 1: The P-Value Fallacy. Ann Intern Med 1999;130:995-1004.

3. Goodman SN. Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. Ann Intern Med 1999;130:1005-1013.

4. Braitman LE. Confidence Intervals Extract Clinically Useful Information from Data. Ann Intern Med 1988;108:296-298.

5. Streiner N. *PDQ Statisitics*. Toronto, B.C. Decker, Inc. 1986

6. Glantz SA. *Primer of Biostatistics*. Third edition. New York, McGraw-Hill 1992.