How to Prepare for Journal Club / EBM conference (Friday at Noon).

You can approach this conference in one of two ways:
1) Using the Evidence Cycle
2) Journal Club

When you use the evidence cycle, you start with a clinical scenario, real or imagined, and then go to the evidence to find an answer, which you then take back to the patient to guide therapy.

When you use the journal club style, you start with the evidence, in the form of a journal article or review, which will then be relevant to you or your colleagues at some point in the future for patient management.

The process for each will be described below, and apply not just to Friday conference but also to practical patient care!

1. Using the Evidence Cycle
   In which you use the “5 A” approach popularized by Duke University, where you start with a real clinical problem you have encountered on the wards or in clinic, and use the evidence to help guide that patient’s management.

   The 5 A’s of the evidence cycle:
   - **Assess**: Start with the clinical encounter / problem
   - **Ask**: Construct your relevant clinical question (PICOT)
   - **Acquire**: Select the best resource(s); i.e. use the literature to answer your question
   - **Appraise**: Critically evaluate the resource(s) you have found (or let ACPJC do it for you!)
   - **Apply**: Take the evidence back to the bedside, apply it to your patient

   **Assess**: This can take several different forms, such as a real patient, or a hypothetical patient, or a group of patients (such as “diabetics”). For example, Dr. Andy Bomback admitted a real patient to Med B with the diagnosis of acute lupus nephritis. He wondered whether to use IV cyclophosphamide or oral mycophenolate mofetil as the initial treatment.

   **Ask**: He asked the question in PICOT fashion:
   - **Patient population**: P: in female patients with acute lupus nephritis
   - **Intervention**: I: does oral mycophenolate mofetil
   - **Comparison**: C: compared to IV cyclophosphamide
   - **Outcome**: O: reduce mortality (or induce more remission, or prevent the need for dialysis, or whatever you want the outcome to be)?
   - **Type of study**: T: therapy

   Another PICOT question he concurrently asked:
   - **Patient population**: P: in patients with acute lupus nephritis
   - **Intervention**: I: is IV cyclophosphamide
   - **Comparison**: C: compared to oral mycophenolate mofetil
**O**: more toxic (less cytopenia, alopecia, infertility etc.)?

**T**: harm

**Assess**: Let’s say Dr. Sule Khawaja has a 50 year old male clinic patient who is terrified of colonoscopy. The patient (and Dr. Khawaja) wonder if virtual colonoscopy is as effective as colonoscopy for screening, without the risk of colon perforation.

**Ask**:

**P**: in patients at average risk for colon cancer  
**I**: is CT colonography  
**C**: compared to colonoscopy  
**O**: as sensitive for detecting colon polyps (or colon cancer, or whatever)?  
**T**: diagnosis

Let’s say the same patient’s father recently had an ischemic stroke, and has partially recovered. The patient thinks his father can live alone for now, but fears that another stroke would be devastating or fatal. He asks you what the risk is of his father having another stroke and dying.

**Ask**:

**P**: in patients with ischemic strokes  
**I**: over time (say the first year after ischemic stroke)  
**C**: compared to patients who have never had strokes  
**O**: what is the risk of a second stroke (or of dying, or whatever)?  
**T**: prognosis

Let’s say Dr Kelly Mitchell has a patient who recently had an adenomatous polyp removed on screening colonoscopy. He is interested in any therapy that could potentially reduce his risk of future adenomatous polyps. Dr Mitchell has heard that COX-2 selective NSAIDs have been shown to reduce rates of colon adenomas.

**Ask**:

**P**: in patients with colonic adenomas  
**I**: does celecoxib  
**C**: compared to placebo  
**O**: reduce the rate of recurrent colonic adenomas?  
**T**: therapy

The same patient shouldn’t take the potential therapy if it is associated with increased risk, and Dr Mitchell recalls that Vioxx was pulled from the market for safety reasons.

**P**: in patients with colonic adenomas  
**I**: does celecoxib  
**C**: compared to placebo  
**O**: cause increased rates of adverse cardiovascular outcomes?  
**T**: harm

**Acquire**: It’s now time to go searching for the evidence. If you are using EBM as a practical tool in clinic or on the wards - where time is critically important - and you find an ACPJC or Cochrane review, stop – they’ve already done the work for you. They are the experts. Don’t waste anymore time reading or appraising the original paper(s) unless you want to. Feel confident using what they give you to help manage your patient.
For example, ACPJC has reviewed the following NEJM paper concerning lupus nephritis:
¬ Lupus Nephritis, ACPJC Review

ACPJC has also reviewed the entire topic of CT colonography:
¬ CT colonography, ACPJC Review

ACPJC has reviewed the broad topic of prognosis after stroke, and has also reviewed this pertinent paper:
¬ Stroke prognosis, ACPJC Review

ACPJC has reviewed a paper on cardiovascular risk in patients taking celecoxib as colonic adenoma chemoprevention:
¬ Celecoxib risk, ACPJC Review

For the purpose of Friday conference, however, all of the questions above are going to need to be found on PubMed (MEDLINE) in the form of single (or a few) papers, which you will need to appraise. Learning to efficiently and effectively search on PubMed is an acquired skill. A great resource is the librarians at the Health Sciences Library. They routinely perform PubMed searches, and are trained in helping physicians and scientists search. Start with Kate_McGraw@unc.edu, phone 966-1855.

PubMed searching reveals the following papers:
1. Mycophenolate mofetil or intravenous cyclophosphamide for lupus nephritis
   PubMed ID: 16306519

2. Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults
   PubMed ID: 14657426

   Stroke. 1993;24:796-800.
   PubMed ID: 8506550

   2006;355:873-84.
   PubMed ID: 15713944

Appraise: This is typically done in the form of a critically appraised topic (CAT) sheet. Each of the 4 types of papers – therapy, harm, diagnosis, and prognosis – is critically appraised in a slightly different fashion. The standard format for critically appraising therapy, harm, diagnosis and prognosis papers comes from Gordon Guyatt’s “User’s Guide to the Medical Literature”, which is basically the Bible of EBM and is an
excellent, readable resource which we have copies of in the Chief’s Office. Consider buying this book! Links to some useful critical appraisal resources are presented here; please skip down to the Journal Club section below for more detail.

→ You can use CAT worksheets from McMaster University to help critically appraise your paper: see Journal Club section below

→ There are also several examples of CAT sheets that former UNC residents prepared at the following link from the housestaff homepage: UNC CAT Sheets

→ Examples of CAT (critical appraisal) sheets for two of the papers mentioned above:
  Therapy paper:
  CAT Sheet: Mycophenolate mofetil or intravenous cyclophosphamide for lupus nephritis

  Diagnosis paper:
  CAT Sheet: Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults

Apply: Take what you have learned from reviewing the evidence back to the bedside. Using evidence based medicine should be a practical, rewarding, and even fun endeavor. It is our obligation to our patients to know “why we do what we do”. In some (many?) cases you will find that there is no good evidence in support of commonly employed treatment or diagnostic strategies.

Applying the Number Needed to Treat (NNT) to therapeutic decision making
Applying the Number Needed to Harm (NNH) to potentially harmful therapies
Applying likelihood ratios to diagnostic testing

2. Journal Club
In which you critically review a paper that has recently been published, usually in a leading journal such as NEJM, Annals, JAMA, Lancet, BJM, Archives. Alternatively, you can select a “classic” or “seminal” paper, possibly published many years ago, that has greatly influenced or dictated the way we practice internal medicine. The focus is on the CAT (critically appraised topic) sheet. Typically one of four types of papers is selected: “therapy”, “harm”, “diagnosis”, or “prognosis”.

Therapy
Sample CAT worksheet for a therapy paper, from McMaster University (go ahead and download the paper version, which has answers and sample calculations): Therapy Paper Worksheet
Key concepts:

General.
Therapy papers are not exclusively evaluations of single drugs; for example, treating blood pressure in diabetics is “therapy”. Likewise, drugs may not be involved at all; for example, banding of varices for acute GI bleeding in cirrhotics is therapy.

Are the results valid?

→ Validity of a study is simply closeness to the truth; ask yourself: “Can I believe the results?”
→ Validity is not “all or none”; for any given study, your needle will end up pointing somewhere between 0% and 100% valid. It’s not a yes or no question.
→ The more bias that is introduced into a study, the less valid (less close to the truth) the study becomes.
→ Ways to limit bias:
  ♠ Randomize patients to treatment and control groups. This accounts for both known and unknown prognostic factors at baseline.
  Pearl: If randomization occurs, then p values are not necessary in the traditional “table 1” baseline comparison of the two groups. P values measure the probability that differences between the groups occurred due to chance (and if they were randomized, then we know any differences are from chance, thus rendering p values unhelpful!)
  ♠ Perform an intention to treat analysis. That is, analyze patients based on the group they were initially randomized to, not the one they end up in. This preserves the initial randomization by keeping known and unknown prognostic factors equally distributed between the two groups throughout the study.
  ♥ Blind patients to which group they are in. Otherwise, those who know they are in the treatment arm may report better outcomes because of the placebo effect (more important for subjective outcomes, such as quality of life scores).
  ♦ Blind clinicians to the group allocation. Clinicians may consciously or unconsciously treat the two groups differently if they are not blinded.
  ♥ Blind those assessing outcomes to the group allocation. If the outcome is all cause mortality, this blinding is less important; but if the outcome is at all subjective (what was the cause of death?), this blinding becomes very important.
  ♦ The term “double blind” is useless; you want to know exactly who were and were not blinded to the group allocation (Patients? Clinicians? Outcomes assessors?)
→ Losing a great proportion of patients to follow up may compromise validity.
  Pearl: Do a worst case analysis (for example; assume all patients lost in the treatment group died, and all those lost in the control group lived) and see if it significantly changes the stated result of the trial. If so, validity has been compromised.

What are the results?

→ Some experts argue that the most important outcome is always all-cause mortality. However, practically, we are often interested in other outcomes such as first ever heart
attack (primary prevention), repeat heart attack (secondary prevention), stroke, progression of renal disease, quality of life, days in the hospital, and on and on.

→ Beware of composite endpoints! This is becoming epidemic in cardiology (and other) trials. A composite endpoint of death, MI, and repeat catheterization is not useful if a significant difference is driven entirely by decreased rate of repeat catheterization!

→ How large is the treatment effect? This is commonly measured by relative risk and its corollary, relative risk reduction. Why? Because these are bigger numbers than absolute risk and absolute risk reduction, and thus “look better”.

→ Keep in mind that if the treatment arm has a 0.01 (1%) absolute risk of developing kuru, and the control group has a 0.03 (3%) absolute risk of developing kuru, then the relative risk reduction is a whopping 66%! (Whereas the absolute risk reduction is 0.02, or 2%) The bigger number almost always ends up being the one reported: Our drug decreases the risk of kuru by 66%! vs. Our drug decreases the risk of kuru by 2%!

Pearl: If the type of risk reduction is not specified, always assume it is the relative risk reduction being reported.

→ Formulae:

<table>
<thead>
<tr>
<th></th>
<th>Died</th>
<th>Survived</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin</td>
<td>40</td>
<td>560</td>
</tr>
<tr>
<td>Placebo</td>
<td>60</td>
<td>540</td>
</tr>
</tbody>
</table>

CER = Control Event Rate
EER = Experimental Event Rate

CER = c / (c+d) = 60 / (60 + 540) = 0.1 = 10%
EER = a / (a+b) = 40 / (40 + 560) = 0.066 = 6.66%
RR = Relative Risk of dying = EER / CER = 0.066 / 0.1 = 0.66 = 66%
RRR = Relative Risk Reduction from ASA = 1 – RR = 1 – 0.66 = 0.34 = 34%
ARR = Absolute Risk Reduction = CER – EER = 0.1 – 0.066 = 0.034 = 3.4%\nNNT = Number Needed to Treat = 1/ARR = 1/0.034 = 29

Relative Odds (Odds Ratio) = (a/b)/(c/d) = (ad)/(cb) = 0.64 = 64% *
*(see below in the Harm section for further discussion of odds and risk)

→ How precise is the estimate of the treatment effect? The true risk reduction can never be known. The best we can do is come up with a point estimate. For example, consider what would happen if you conducted a trial of ASA vs. placebo, and came up with an ARR for death of 3.4%. Then, let’s say a month later you repeated the study with a different group of patients. Do you think the ARR for death would once again be 3.4%? What if you repeated the study 100 times, would you get 3.4% each time? Unlikely!

→ We then place that point estimate within a neighborhood in which the true value is likely to fall: the confidence intervals. The larger the study, the narrower the confidence intervals will be. In the above example, the 95% confidence interval might be 1.2%-6.9%, with the point estimate being 3.4%.

→ We feel comfortable that ASA is actually preventing death when our 95% confidence intervals do not cross 0.0%. If that happens, then we can’t be confident that ASA is actually preventing death at all!

♦ Using the above example, you can state: “I am 95% sure that the true treatment benefit of ASA lies between 1.2% and 6.9% ARR”.
The more confident you want to be, the broader the C.I. range will be. For example, if you wanted to be 98% certain that the true value lies between your confidence intervals, the range might be 0.07% to 20.3%.

What is a p value? Simply, the probability that the observed difference is due to chance. If the p value for ASA vs. placebo is 0.04, then we can make the statement: “I am 96% sure that ASA is preventing death compared to placebo. But, I acknowledge that there is a 4% possibility that the difference I see is due to chance, and that ASA is no better than placebo”.

Pearl: Assigning statistical significance to p values of <0.05 and C.I.’s of 95% is purely arbitrary, but widely accepted. Ask yourself, is a p value of 0.06 not helpful?

Harm
Sample CAT worksheet for a harm paper, from McMaster University (go ahead and download the paper version, which has answers and sample calculations):
Harm Paper Worksheet

Key Concepts:
General.
It is not ethical to prospectively randomize a group of patients to a potentially harmful intervention for the purpose of demonstrating harm versus a placebo group. How, then, are potentially harmful therapies or exposures recognized? Several ways: RCTs, cohort studies, case control studies, and case reports.

Are the results valid?
Consider whether there was similarity in all known determinants of outcome. This is more likely to have occurred in an RCT, and less likely in cohort and case-control studies.

♦ If not, were differences adjusted for?

RCTs (least biased, least sensitivity)
♦ The most unbiased type of “harm” study is a randomized control trial designed to show the benefit of a particular therapy (i.e. a “therapy” paper), in which harm is incidentally found; or in which harm was expected, but unexpectedly outweighed benefit.
♥ Take for example the Women’s Health Initiative. This was a randomized controlled trial which was supposed to show the benefits of estrogen/progestin in menopausal women for primary prevention of coronary heart disease. Instead, this arm of the trial was stopped early, as it surprisingly found a significant increase in adverse cardiovascular outcomes in the estrogen/progestin group.
♦ Likewise, a trial looking at Celebrex as chemoprevention of colonic adenomas was stopped early when significantly increased cardiovascular risk was demonstrated in the Celebrex group (despite Celebrex being very effective in prevention of adenomas). These are just two examples of how “therapy” papers may also be “harm” papers.
♦ One potential drawback of using RCTs as harm studies occurs when the harmful side effect occurs infrequently, such as clopidogrel-induced TTP. This
occurs so rarely that it has not been detected even in large RCTs (CAPRIE, CURE), but has been reported in case reports and case series.

**Cohort Studies** (more biased, more sensitive)

♦ A cohort study involves identifying two groups of people – exposed and non-exposed – (say smokers and non-smokers), who are otherwise similar, and following them over time to determine rates of developing an adverse outcome (say lung cancer).

♥ One benefit of a cohort study is that infrequent outcomes can be more easily detected than with an RCT.

- For example: if the rate of GI hemorrhage in NSAID users is 1.5 per 1000 person-years, and the rate of GI hemorrhage in non-NSAID users is 1.0 per 1000 person-years – it would take an RCT of 75,000 patients to detect this small, but real, harmful effect. A cohort study could draw from a large database (e.g. Medicare patients, or VA patients) and easily evaluate 75,000 or more patients.

♦ One drawback of a cohort study is the potential for confounding variables.

- For example, what if the real reason for more GI bleeding in the NSAID cohort is advanced age and not the NSAIDs themselves (older people have more aches and pains, and thus take more NSAIDs)?
- Or, what if people with arthritis have higher rates of GI hemorrhage independent of NSAID use?
- The best we can hope for is that investigators matched baseline variables as well as possible, and that differences that do exist are adjusted for statistically.

Pearl: A cohort trial will always inherently have more bias than a randomized trial.

**Case Control Studies** (most biased, most sensitive)

As mentioned above, RCTs will limit bias between the exposed and non-exposed groups by the process of randomization itself, with the drawback of decreased sensitivity for detecting infrequently occurring harmful side effects. Cohort studies have the capability of evaluating huge numbers of patients (entire databases), and thus detect more of the infrequently occurring harmful side effects, with the drawback of bias that cannot be completely eliminated.

♣ Case control studies are 100% sensitive, because they start with a group of patients who all have suffered the adverse side effect (the “case” population). This case population is then retrospectively matched with a group of people of similar age, sex, and other baseline characteristics but who have not suffered the adverse side effect (the “control” population). Thus, working backwards, an attempt is made to determine what the harmful exposure may have been.

♦ This sort of analysis is prone to tremendous amounts of bias as it is retrospective and obviously not randomized, plus there is a natural tendency towards “recall bias” and “interviewer bias” [when the case population better remembers (recall) or is more aggressively asked about (interviewer) - previous exposures than the control population].

Were exposed patients equally likely to be identified in the two groups? This refers to recall and interviewer bias.
Were outcomes measured in the same way in the groups being compared? This refers to surveillance bias. For example, investigators may search more diligently for skin cancer in the lifeguard group as opposed to the librarian group, thus finding more melanomas and finding them at earlier stages. A conclusion that lifeguards have more melanoma than librarians may be falsely influenced by surveillance bias.

Was follow-up sufficiently complete? Again, do a worst case scenario (assume all those lost in the exposed group lived, and all those in the non-exposed group died), and if the results are not significantly changed, then validity was not compromised.

**What are the results?**

How strong is the association between exposure and outcome?

- For cohort studies, the Relative Risk can be calculated
- For case control studies, the Relative Odds (Odds Ratio) must be calculated

Pearl: When the event rate is low (low prevalence), the odds ratio and relative risk are very similar – close enough to be basically the same. As the event rate increases, however, they separate.

For example, if the odds are 1:1 (1.0), then the risk is 1/2 (.50)

And, if the odds are 1:5 (.20), then the risk is 1/6 (.17).

And, if the odds are 1:9 (.11), then the risk is 1/10 (.10)

And, if the odds are 1:99 (.01), then the risk is 1/100 (.01)

**Formulae:**

<table>
<thead>
<tr>
<th></th>
<th>Lung CA</th>
<th>No CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>80 (a)</td>
<td>920 (b)</td>
</tr>
<tr>
<td>Non smoker</td>
<td>20 (c)</td>
<td>980 (d)</td>
</tr>
</tbody>
</table>

RR (if cohort study) = \(\frac{a/(a+b)}{c/(c+d)}\) = \(\frac{80/1000}{20/1000}\) = 4

OR (if case control) = \(\frac{(a/b)/(c/d)}\) = \(\frac{80/920}{20/980}\) = 4.26

Absolute risk increase = EER - CER = .08 - .02 = .06 = 6%

Number Needed to Harm (NNH) = 1/.06 = 16 smokers

How precise is the estimate of risk? Confidence intervals. The wider the confidence intervals, the less precise the point estimate of risk.

**Diagnosis**

Sample CAT worksheet for a diagnosis paper, from McMaster University (go ahead and download the paper version, which has answers and sample calculations):

Diagnosis Paper Worksheet

**Key Concepts:**

**General**

The critical concepts in diagnostic testing are:

1. Diagnostic certainty and treatment threshold
2. Morbidity / mortality risk of the illness
3. Morbidity / mortality risk of the treatment
4. Pre-test probability of disease
5. Positive and negative likelihood ratios
6. Post-test probability of disease
7. Tests are not always dichotomous variables

- Diagnostic certainty is closely related to treatment threshold
  ♦ If the illness carries low morbidity or mortality risk (viral URI), and the proposed treatment is also benign (fluids or acetaminophen), then the diagnostic certainty does not need to be very high for action to be taken, and thus no diagnostic test needs to be performed.
  ♦ If the illness carries high morbidity or mortality risk (for example PE), the treatment threshold may not be very high. You probably would treat for PE if there was a 15% chance that PE is the correct diagnosis.
  ♦ If the treatment is not benign (bone marrow transplant, thrombolytics), then the diagnostic certainty needs to be fairly high before initiating treatment, and a diagnostic test needs to be performed.

- Pre-test probability of disease
  ♥ If the prevalence of disease in a certain patient population is known, it can be used as the pre-test probability.
  ♦ If a clinical prediction rule exists, it may be used to determine the pre-test probability (e.g. Wells criteria for DVT or PE).
    Example: Link to CPRs
  ♦ Your clinical acumen and experience may be used to determine the pre-test probability.

- Likelihood ratios. (link to LR nomogram)
  ♦ For any given test, each result has its own likelihood ratio. It can be argued that there is no such thing as a “positive” or “negative” test result; tests only make you either more or less certain of disease than you were before the test.
    - The LR is the numerical value for how much more or less certain you are of disease after obtaining a test result. If the LR is >1.0, then you are more certain of disease than you were before the test was done. If the LR is <1.0, then you are less certain of disease than you were before the test.
  ♥ Consider the diagnosis of iron deficiency. Let’s say a serum ferritin level is performed on 260 elderly patients who also receive bone marrow biopsies (gold standard for diagnosing iron deficiency). Every single ferritin result (from 1 ng/ml to 2000 ng/ml) has a likelihood ratio associated with it. At some point, say around 40 ng/ml, the likelihood ratio equals one. At that point – where the likelihood ratio is 1 – as many people with that particular test result (ferritin = 40) have the disease (iron deficiency) as do not.
    Pearl: A likelihood ratio of 1 is the only test result which is completely unhelpful to you; all other values will either make you more or less certain of disease than you were before testing.
    Pearl: A strong + LR is typically 10 or greater. A strong – LR is typically 0.10 or lower.

- Test results are not always dichotomous variables.
  ♦ The UNC lab uses 3 ng/ml as the cutoff for a “normal” ferritin. Let’s examine why assigning “normal” or “abnormal” to lab results is often misleading.
    Example: Let’s say the pre-test probability of iron deficiency in your elderly patient is 20%. Oral iron is somewhat unpleasant to take – and iron overload can be harmful - so you do want to perform a test to help
decide whether the patient is truly iron deficient before starting iron replacement therapy. The serum ferritin value returns as 8 ng/ml, which the lab reports as “normal”. Ferritin of 8 has a + LR for iron deficiency of around 40. Thus, the probability of iron deficiency after performing this test is 90%! The test moved you from 20% probability of disease to 90% probability of disease.

♣ This logic can be applied to almost all tests: 3 mm of ST elevation after 2 minutes of exercise on a treadmill is a “positive” test result, as is failure of HR to recover after 12 minutes of exercise. Despite both being “positive” test results, they would clearly have different positive likelihood ratios for diagnosis of CAD.

➔ Link to Likelihood Ratios for common physical exam findings based on the JAMA rational clinical examination series
➔ Link to Likelihood Ratios for commonly ordered tests, developed by Mike Pignone MD, UNC Dept of General Medicine

Are the results valid?
➔ Did clinicians face diagnostic uncertainty? We don’t need a test that rules out MI in 20 year old women with rib pain; likewise we don’t need a test that tells us that grouped painful vesicles in an elderly man are herpes zoster. We need help in the grey area in-between these extremes.

♦ Sometimes pre-test probability of disease is so low that it does not even reach your test threshold; that is to say, if the test result were positive, you wouldn’t believe it (you would consider it a false positive).

♥ Likewise, there are situations where you are so certain of disease that you would not believe a negative test result (you would consider it a false negative).

♠ In these cases – where you are already below your test threshold or above your treatment threshold – a test is not helpful, because in these situations there is no diagnostic uncertainty.

➔ The way this applies to the validity of your Journal Club paper is summed up by asking the following questions: did the investigators test appropriate patients, and are my patients similar to the patients in the study?

➔ The Gold Standard: the accuracy of a test is determined by comparison with the reference standard (representing “the truth”). No gold standard is perfect (not even biopsy or autopsy), so you have to ask yourself whether you accept the investigators’ gold standard.

Pearl: The gold standard is unacceptable if it includes the test being studied.

➔ If you accept the gold standard, next assess whether the test in question and the gold standard test were interpreted blindly (not being blinded introduces bias; for example, we can all hear the murmur after we’ve seen the echo; we can all find the spot on CXR after we’ve seen the CT).

➔ Long term follow up is usually an acceptable gold standard, but look to see if the investigators performing the follow up were blinded to test results.

What are the results?
➔ The 2x2 table. You use a two by two table when results are reported in dichotomous fashion. You can also use 2x3, 2x4 or 2 by whatever for non-dichotomous results.
“The truth goes on top”. Thus, the gold standard (representing truth), goes on top of your table.

Consider the previous example of the diagnosis of iron deficiency anemia:

- **The test**: ferritin
- **The gold standard**: bone marrow biopsy

### 2x2 table

<table>
<thead>
<tr>
<th></th>
<th>BMBx positive</th>
<th>BMBx negative</th>
<th>+ LR</th>
<th>- LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ferritin &lt;19</td>
<td>47 (a) (TP)</td>
<td>2 (b) (FP)</td>
<td>41.5</td>
<td></td>
</tr>
<tr>
<td>Ferritin ≥ 19</td>
<td>38 (c) (FN)</td>
<td>148 (d) (TN)</td>
<td></td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>150</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Formulae

- **LR+** = \( \frac{TP}{TP + FN} \) / \( \frac{FP}{FP + TN} \)
- **LR-** = \( \frac{FN}{FN + TP} \) / \( \frac{TN}{TN + FP} \)
- **LR+** = \( \frac{a}{a+c} \) / \( \frac{b}{b+d} \)
- **LR-** = \( \frac{c}{a+c} \) / \( \frac{d}{d+b} \)
- **Sens** = TP / (TP + FN)
- **Spec** = TN / (TN + FP)
- **PPV** = TP / (TP + FP)
- **NPV** = TN / (TN + FN)

### 2x4 table

<table>
<thead>
<tr>
<th></th>
<th>BMBx positive</th>
<th>BMBx negative</th>
<th>+ LR</th>
<th>- LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ferritin &lt;19</td>
<td>47</td>
<td>2</td>
<td>41.5</td>
<td></td>
</tr>
<tr>
<td>19-45</td>
<td>23</td>
<td>13</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>46-100</td>
<td>7</td>
<td>27</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>&gt;100</td>
<td>8</td>
<td>108</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>150</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Likelihood ratio in words: a serum ferritin value of less than 19 ng/ml is 41 times more likely to occur in a patient with iron deficiency than in an iron replete patient.

### Converting pre-test probability to post-test probability

- **Step 1.** Convert pre test probability to odds
- **Step 2.** Multiple odds by LR
- **Step 3.** Convert post test odds back to probability

**Example:**
- Pre-test probability of iron deficiency: 20%, or 20/100, or 1/5
- Convert probability to odds: odds = 20:80 (or, 1/4)
- Ferritin = 15 ng/ml, LR = 41 (see table above)
- Multiply the odds by the LR: \( (1/4) \times 41 = 10 \)
- Post test odds: 10:1 (10/1)
- Convert to probability = 10/11, or 91%
Why bother converting probability to odds?

What if I said I watched the news last night, and they said the probability of rain on Tuesday is 60%. Today I hear the weatherman say it’s now twice as likely to rain on Tuesday. So what is the probability of rain on Tuesday now? Is it 120%? No! You can’t multiply probabilities, you can only multiply odds!

Initial probability of rain: 60%, or 60/100
Initial odds: 60:40, or 3:2, or 3/2
LR = 2 (twice as likely)
Multiple odds by LR: 3/2 x 2 = 6/2 = 3
Revised odds = 3:1, or 3/1
Revised probability of rain on Tuesday = 3/4, or 75%

Why is the positive predictive value less helpful than the positive likelihood ratio? Because the PPV is entirely dependent on the underlying prevalence of disease in the population being studied, something we don’t often know.

Consider the example of using the ANA test for diagnosis of lupus. No matter what patient you are testing, the LR+ for the ANA test is 19 (hypothetical), and does not change.

♣ So, if you perform the ANA on a 70 year old man with joint aches, the LR+ is 19; and if you use perform the ANA on a 40 year old female with joint pain, photosensitive rash, and pleurisy, the LR+ is 19.
♦ However, in the first patient the PPV for ANA is 1.9%; and in the second patient, the PPV for ANA is 95%. A huge difference in the property of the ANA test from patient to patient!
♥ What is the difference between the 70 year old with joint aches and the 40 year old with joint aches, photosensitive rash and pleurisy? The difference is the prevalence of lupus in these two very different patient populations. In 70 year olds with joint aches, the prevalence of lupus is less than 1% (say 0.10%). In 40 year olds with three of the most common lupus symptoms, the prevalence of lupus is much higher (say 50%).
♠ Link here for ANA worksheet

While the LR doesn’t change from patient to patient, the predictive values may vary widely. Thus PPVs are largely unhelpful to us, while LRs are very useful.

For conditions like iron deficiency and hypothyroidism, there may be multiple diagnostic tests available. For iron deficiency, there are the ferritin, mean cell volume (MCV), transferrin saturation (T Sat), red cell volume distribution, and others. Likewise for hypothyroidism there are the TSH, total T4 level, free T4 level, thyroid binding index, and others. Why have ferritin and TSH been chosen as the tests of choice?

The receiver operator characteristic curve (ROC curve) plots the TP rate on the y axis and the FP rate on the x axis. The plot with the largest area under the curve is the best test. The point at which the slope is such that one sees the most TPs for the fewest FPs is often chosen as the test “cut-off” between normal and abnormal. Every point on the curve represents a positive likelihood ratio (TP rate over FP rate).
Prognosis
Sample CAT worksheet for a prognosis paper, from McMaster University (go ahead and download the paper version, which has answers and sample calculations):
Prognosis paper Worksheet

Key Concepts:
General