EVIDENCE BASED MEDICINE
Bill Lefkowitz

DISCLOSURE
I have no relevant financial relationships with commercial interests.

ETHICS
• "VIII: A physician shall, while caring for a patient, regard responsibility to the patient as paramount" - AMA, Principles of Medical Ethics, June 2002 revision
  • Duty to be the best provider you can be
  • Duty to practice "Evidence Based Medicine"
    • which is the best kind of medicine
OPENING QUESTIONS

- What is evidence based medicine?
  - No fighting please.

- If the result of a study of a novel therapy vs. standard therapy has a p-value of 0.20 (or a 95% CI of 0.8-2.0 if you prefer), does that mean both therapies are equally effective?

- Which represents a higher quality of evidence?
  - Case series or cohort study?
  - RCT or expert opinion?
  - n-of-1 trial or meta-analysis?

OVERVIEW

- Define Evidence Based Medicine
- Describe the tools of EBM
- Discuss the magic of the RCT
  - Our inherent quasi-Bayesian nature
- Discuss the limits of evidence based medicine applied to direct patient care
- The LEAP model of evidence based medicine
- Ethical implications for the integration of evidence based medicine into clinical practice

WHAT IS EVIDENCE BASED MEDICINE?

- The old paradigm
  - Understanding pathophysiology is sufficient
  - Guidelines based on experience
  - Expertise based on experience
- The new paradigm
  - Intuition may be misleading
  - Pathophysiologic understanding is insufficient when tests of accuracy, efficacy exist
  - Grateful Med, library trip, photocopy, 30 minutes, $2.68
WHO DEVELOPED THE IDEA?

- Most of working group made of MD, MSc from the departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster Univ. Ontario
  - Gordon Guyatt (chair) coined the term in 1990
  - David Sackett (founded dept of epi/bio)
- Primarily people who understood the mechanics of the scientific process (hypothesis testing) in an environment where proof took the form of experts recounting personal experiences and observations
  - By then research methods had drifted into medicine, but were underappreciated, questionable applied
  - As opposed to the pre 1980s
- So, “know-it-alls” telling us to “de-emphasize” the current experts

THE TRANSITION

- The push-back
  - EBM ignores clinical experience and clinical intuition
  - Understanding pathophysiology plays no part in EBM
  - EBM ignores basic clinical training
  - “De-emphasize” the old standards
    - Expert clinician opinion based on experience, and clinical intuition
  - “Emphasize” the new, when available
    - Epidemiology and research, suspicion of biased experience

WHY DID IT CATCH ON?

- Focused on educators (spoke their language)
  - Metrics! Yay!
  - Curriculum, tools, series of “how to” articles in JAMA
- Focused on the young educated
- MOST CRITICAL FACTOR: give the young a tool that they can use as a short-cut to the “10,000 hours” of focused clinical practice that are needed to be considered an expert, and empower them to consider that literature derived knowledge was of better quality than clinical expertise
  - Takes very little reading to become an “expert” on a particular condition
  - Much quicker return on investment of education
  - If evidence existed, a librarian could provide better care for a patient than a physician
MISSED ADAPTATION

- EBM developers understood the limits of literature and eventually shifted language from “de-emphasizing” to “integrating.”
  - Their disciples didn’t necessarily catch the acknowledgement of the importance of clinical experience
  - Intuition may be misleading
    - So might poorly performed/interpreted studies (esp. retrospective, regression)
  - Lorber, Wiswell
- As a consequence of that under-understanding, the pendulum over-swung
- Despite the fact that EBM always meant that one should use the best available evidence to make clinical decisions, the (not entirely) unintended message was that the only evidence worth a damn was derived from research methods
  - And the RCT the best quality evidence that research methods have to offer

THE TOOLS OF EVIDENCE BASED MEDICINE

- THERAPY from “Levels of Evidence 1” (1998)
  - 1: Randomized Controlled Trial and reviews of RCTs
  - 2: Cohort studies and reviews of cohort studies
  - 3: Case-control studies and reviews of case control
  - 4: Case reports and case series
  - 5: Expert opinion without explicit critical literature appraisal, or based on bench/phys
    - [Paradox? ... what kind of expert now doesn’t know the literature?]


Systematic Reviews and Meta Analysis
Randomized Controlled Trials
Cohort Studies
Case Control Studies
Case Series and Reports
Ideas, Editorials and Opinions

Strength of Evidence
<table>
<thead>
<tr>
<th>Levels of Evidence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: n-of-1 trials or systematic reviews of RCTs</td>
<td></td>
</tr>
<tr>
<td>2: randomized controlled trial or observation study with dramatic effect</td>
<td></td>
</tr>
<tr>
<td>3: non-randomized control or cohort studies</td>
<td></td>
</tr>
<tr>
<td>4: case series, case control and historical control</td>
<td></td>
</tr>
<tr>
<td>5: mechanism based reasoning</td>
<td></td>
</tr>
</tbody>
</table>

THE MAGIC OF THE RCT

- What's the big deal about the RCT?
- It's a shortcut to gaining vast experience and contemplating your tummy button
- It weeds out the biases if set up properly
- Name something we "know" without the benefit of RCT
  - Smoking/cancer
  - EBM/NEC/intelligence
- So, how does it work?

THE CONCEPT

- Consider two groups of patients who are alike in all but one way
- We pick some defining characteristic of the samples to measure (dichotomous) and compare (means, ranks, variances, etc)
- What are the chances that I could pick two samples whose means are that far (or farther) apart from each other from the same population?
- If we want to use frequentist statistics, then BEFORE we start the experiment we have to calculate the probability of every possible outcome, so we know how to interpret the result
- What if we did 100 trials? (p-values)
  - If I flip a coin five times and it lands heads each time, what are the chances that it happened?
    - Assuming that I planned on 5 flips and used a balanced coin, 3.125% (0.03125)
- What is the purpose of the "random sample?"
THE RANDOM SAMPLE

- So that we can calculate the probability of selecting a sample whose mean is at least "x" far away from the population’s mean
- Can only be calculated if THAT SAMPLE IS SELECTED RANDOMLY
- Do we really use random sampled?
  - No. So filling a technical standpoint applying the study outcome to the conceptual population from which we pretend we drew is kind of a leap of faith
  - What kind of sample is it?
    - We “test” it (via: “table 1”)
  - We divide them randomly because a (pretend) random sample divided randomly is two random samples
- What happens if we didn’t select (or at least randomize) “randomly”
  - None of the statistical tests are valid if the sample isn’t random (despite the “p” you got)
  - Too good to be true (type 3 error)

THE P-VALUE

- So what exactly is the p-value?
  - The probability of getting a result as extreme, or more extreme, as the one we got IF THE NULL HYPOTHESIS IS TRUE
  - If we select two samples at random from a single population, what are the chances they would be this far (or farther) apart IF THE DIFFERENCE BETWEEN THEM WAS SUPPOSED TO BE ZERO?
  - Why do we use 0.05?

- So if I conduct a study and I get a p-value of <0.05, does that mean the null hypothesis is (probably) false (and I can accept the alternative)?

THE P-FALLACY

- No! Strangely. But not strange to a frequentist...
  - This method is designed to turn out the true result in the “long run”
  - If I use a cutoff of 0.05 in the “null hypothesis rule-out approach,” about every 20 times I run the experiment, my two samples should be “too far apart to have been randomly selected from the same population for my comfort” in an experiment in which the null hypothesis is true (type 1 error)
  - Whether it’s the first run, the 12th run or the 126th run we can’t know (random?)

- Even if I have an experiment in which the null hypothesis is false I will occasionally get a non-significant result (type 2 error)
WHAT THE HEC K?

- So... why even bother?
  - Any individual experiment, run once, is subject to this possibility
  - Just run the experiment multiple times ($n$, time)
- Because we are quasi-Bayesians
- How do the frequentist and Bayesian interpretations of the same study differ?
  - Frequentist: how likely is it that I got this result if the truth is $H_0$?
  - After all, the truth IS or ISN’T
  - When I’m done applying the intervention an infinite number of times, the outcome, on average, will be “the truth”
  - Bayesian: what does this particular result tell me about the truth?
  - The outcomes of my study ARE “THE TRUTH”
  - After all, they are what happened
  - Along with everything else I know, how well does it help me predict what will happen next time?

YOU KNOW YOU’RE A QUASI-BAYESIAN IF:

- You’ve ever said “a p-value of <0.05:
  - Is probably a statistical anomaly, it needs more support” when you expected no difference
- You’ve ever said “a p-value of >0.05:
  - Means there is no difference because we did a power analysis and had an 80% chance of finding a difference if it existed” when you expected no difference
- “Is a trend towards significance” when you expected to see a difference
- “Is probably just because the study is underpowered” when you expected to see a difference
- The common thread: a-priori expectations influence the interpretation of the results

“STRENGTH” OF A STUDY

- In reality, no study is undertaken in the total absence of evidence
- If there’s a good back-story, a weak study will suffice, if the results are surprising or contrary to current belief, a strong study is needed
- How does one define a “WEAK” or a “STRONG” study?
- One thinks very differently as a clinician vs. a researcher
  - Clinicians prove themselves right
    - Develop a (clinical) hypothesis and look for supporting evidence
  - Researchers prove themselves wrong, and in proving themselves wrong, prove themselves right
    - Develop a hypothesis, then a (false) hypothesis to try to prove wrong, so that the alternative (the true hypothesis) must be right (if the coin didn’t land on heads, it must have landed on tails)
“STRENGTH” OF A STUDY

- Proving yourself right (weak “proof,” circumstantial, cherry picking evidence)
- Proving yourself wrong (stronger proof, and frankly the only tool that statistical methods offer ... but feels more objective, as does getting a number out of it)

<table>
<thead>
<tr>
<th>Skeptics</th>
<th>Strong study</th>
<th>Weak study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Believers</td>
<td>Believe</td>
<td>Believe</td>
</tr>
<tr>
<td>Skeptics</td>
<td>Believe</td>
<td>Skeptical</td>
</tr>
</tbody>
</table>

THEREFORE: ethically obligated to use strong research methods when committing patients to be research subjects.

RC T IS THE GOLD STANDARD

- EVEN assuming perfect research methodology there are limits to what the RCT can tell us.
- The gold standard for WHAT?
- For evaluating the effect of an intervention on a disease in a particular patient population
  - Population defined by inclusion/exclusion criteria
- And it gives you information about disease without messy confounders
  - Known and unknown confounders are nullified through the use of randomization
  - Disease centric ... very allopathic

EQUIPOISE

- Gear shift ... What is equipoise?
- Uncertainty about whether an intervention will be beneficial
- Ethically required for randomization
- If the preponderance of evidence is that something is harmful, you can’t study it
- If the preponderance of evidence is that something is beneficial, you can’t study it
- The trick, then, is deciding what constitutes a preponderance of evidence
  - Or, who? IRB, OHRP, AHRP? (office/alliance for human research protection)
CONFOUNDERS

- Does everyone benefit in the treatment arm of a successful trial?
  - No. If they were likely to, there wouldn't be enough equipoise to study the intervention.
- Example:
  - The TIPP trial (PDA: 24/50%, Gd 3-4 IVH: 9/13%)
  - The CAP trial (BPD: 36/47%, PDA surgery: 5/13%)
  - The SUPPORT trial (Severe ROP: 9/13%, RIP: 20/16%)
  - The COSS trial (revised) (BPD: 39/45%, RIP: 23/16%)
- BLOB: there was still a high proportion of bad outcomes in the study arm that showed better outcomes.
  - NOT EVERY SUBJECT BENEFITS FROM THE TREATMENT, NOT EVERY SUBJECT IS HARMED BY NOT GETTING THE TREATMENT.

A 750G KID IS BORN... WHAT DO YOU DO?

- Now the fun part... you know the data, how do you apply it?
  - Now that I know the data, what is my ethical practice obligation?

  - Set goals of 85-89% v. 91-95%
  - Do you follow the data?
  - Do you pick a passive-aggressive mid-range? (88-92 wasn't studied and is probably OK)?
  - Do you pick what's "best for the baby"?
    - What exactly is best for the baby? How do you know?
  - Do you rely on sub-group analysis?
    - What's wrong with using sub-group analysis?
      - A-priori or purely post-hoc
      - Texas sharp-shooter fallacy

For evaluating the effect of an intervention on a disease in a particular patient population:

- Really, it tells you the typical response of the intervention on the disease in a generalized group of patients.
- Who takes care of populations?
  - Public Health
  - Clinical practice is not full-on public-health-style management. It's full of individual patients.
THE BABY WITH CONFOUNDERS

• Just because the kid matches the entry criteria, does that mean the successful study intervention represents the best therapy for this individual?

  • NO!
  • For the disease in this population, IN THE ABSENCE OF CONFOUNDERS, certainly it suggests the experimental therapy is better than control (though not necessarily the “best”) as far as the outcome measures are concerned
  • Recall the whole point of RCTs is to “randomize-out” the confounders
  • But who takes care of patients without confounders?
  • And some of those confounders are MUCH stronger and more relevant to the patient than the effect of the intervention

OTHER LIMITATIONS

• What else limits the applicability of RCTs?
  1. Improperly performed RCTs
  2. Improper interpretation of RCTs (often answering questions you weren’t asking)
  3. Questionable measurements (time to full feeding)
  4. Necessity is the mother of invention (If my unit doesn’t have a nec problem….)
  5. Single interventions vs. packages of care (overly simplistic studies)
  6. Alternative, untested therapies (no one tested intranasal decadron drops/BPD)
  7. Modifications of the conceptual population (apply TIPP to 1100g kid)
  8. Modifications of the therapy (0.1mg/kg decadron, not 0.5mg/kg)
  9. Improperly scoped search for evidence
    • (Fundamental shortcoming of the “vision.” wrong question, premature stopping …)
  10. etc.

THE LEAP MODEL

• When looking at the effect of a therapy on a disease in a specific population the RCT/meta-analysis is the best evidence possible
  • Research methods have given us the ability to expand our knowledge about diseases
  • The clinician’s job is to apply the evidence to patient care (this isn’t addressed in the “strength of evidence pyramid”)

This presentation is the intellectual property of the author.
Contact them for permission to reprint and/or distribute.
THE LEAP MODEL

- LEAP: Literature and Experience Applied to Practice
  (alternate: iCRAP: integrated Clinical and Research Applied to Practice)

- The bottom half is the original strength of evidence pyramid, but emphasizes how the weight of evidence related to the understanding of a disease in a particular population grows as the strength of evidence improves
  - (without RCTs the base for the top half is a lot smaller)

The LEAP MODEL

- The top half describes the layers of evidence available to the clinician as they are increasingly able to incorporate more information specific to the individual patient
AREN’T YOU LETTING IN BIASES?

- Of course. The only way to remove bias is to limit study populations to kids with the same (known and unknown!) confounders as your patient.
- Not enough time/resources for such an endeavor...
- Really… It’s expensive.
- All evidence has limits.
- Your job as a clinician is to balance all sorts of evidence and make a final decision.
- Evidence includes everything/ even clinical instinct.
  - Research
  - Lab studies
  - Exam
  - History
  - Personal preferences
  - Appreciation of available resources
  - Proposed mechanisms
  - Etc.
- Each has a weight, and nothing is of no importance.

ETHICS

- "VIII: A physician shall, while caring for a patient, regard responsibility to the patient as paramount" - AMA, Principles of Medical Ethics, June 2001 revision

- Duty to be the best provider you can be
- Duty to practice “Evidence Based Medicine”
  - which is the best kind of medicine
- Duty to understand the use and limits of the literature in evidence based medicine
- Duty to make each patient encounter a learning opportunity
DEEP THOUGHTS

- Research can supplement, but not replace, reason
- Reading the literature can supplement, but not replace, experience
- Principles of evidence based medicine cannot, and should not, supplant classical clinical education

FINAL POINTS

- Research methods can help you better understand a disease
- As a clinician you need to understand your patient
  - who happens to have a disease
  - and confounders
    - and goals
    - and beliefs and values
    - and a particular treatment plan
    - and they sneezed on the medications in your waiting room
  - and they took all the lollipops, so whatever...

FINAL POINT

- Understanding patients means learning CLINICAL/physical medicine
- EVERY patient encounter is a unique learning opportunity
  - Literature (clinical research)? Lumps, experts clinicians split
- How many kids with RDS do you need to see?
  - Challenge your clinical predictions and look for unique features in every kid
  - If you haven’t walked away from each patient encounter with at least a few ponderables or interesting observations/questions, you haven’t taken advantage of the learning opportunity
ANSWERS TO OPENING QUESTIONS

• What is evidence based medicine?
  • Using the best available evidence to make a clinical decision. At issue is what you consider “good evidence.”

• If the result of a study of a novel therapy vs. standard therapy has a p-value of 0.20, does that mean both therapies are equally effective?
  • Frequentist: no … can’t prove equivalance
  • Quasi-Bayesian: sure, if there’s no other evidence that it’s more effective

• Which represents a higher quality of evidence?
  • Case series/cohort study? (“classic” cohort)
  • RCT or expert opinion? (“classic” RCT)
  • n-of-1 trial or meta-analysis? (“classic” equal)

QUESTIONS?

(the end)