



***Evidence Based Medicine:  
Why Good Studies Turn Out  
To Be Wrong***

*Allan Garland, MD, MA*

*Associate Professor of Medicine and Community Health Sciences*

*University of Manitoba*

## *Evidence-Based Medicine (EBM)*

---

- All aspects of medical practice should be determined by the best available evidence
  - not anecdotes, “in my experience”, or “that’s how we do it here”
  - personal experiences are prone to serious biases (e.g. recall bias), and are not science
- EBM stresses evidence from clinical research
  - and places a low value on authority as a rationale
- But not all published studies are created equal

# Reasons Studies May Get the Wrong Answer

- There are LOTS of reasons (GIGO)
- A wide range of issues related to flawed study design/execution
  - numerous kinds of bias -- e.g:
    - ◆ use of historical controls (*Am J Med* 72:233,1982)
    - ◆ poor blinding; poor concealment of randomization (*JAMA* 273:408,1995)
  - use of surrogate or physiologic endpoints:
    - ◆ 3 RCTs show that NO in ARDS improved  $P_{O_2}/F_{I_{O_2}}$   $\Rightarrow$  none showed improved survival (*CCM* 26:15,1998)
    - ◆ original study of nesiritide for severe CHF showed it was better than iv NTG at lowering PCWP (*JAMA* 287:1531,2002)  $\Rightarrow$  turned out to increase risk of renal failure and death (*JAMA* 293:1900,2005)

# Reasons Studies May Get the Wrong Answer

- A wide range of issues related to statistical analysis
  - inappropriate statistical test/methods -- e.g:
    - ◆ categorizing continuous variables (*Stat Med* 25:127,2006)
    - ◆ use of univariate analysis as entry point to multivariable adjustment (*J Clin Epi* 49:907,1996)
    - ◆ failure to account for immortal time bias (*AJRCCM* 173:842,2006)
  - failure to validate the assumptions of the statistical methods (*Ann Intern Med* 118:201,1993)
- ★ BUT -- Even many studies without any obvious such flaws turn out to be wrong ⇒ Because the standard approach to interpreting “the truth” about studies is *fundamentally* weak

## *A Few Examples*

---

- Initial studies that turned out to be wrong:
  - H1-A1 anti-endotoxin antibody and mortality in septic shock (*NEJM* 324:429,1991 vs. *Ann Intern Med.* 121:1,1994)
  - Estrogen-progesterone and CAD in postmenopausal women (*JAMA* 273:199,1995 vs. *JAMA* 288:321,2002)
- Studies supporting current practices that may well turn out to be wrong:
  - Tight glucose control (4.5-6 mM) in the ICU (*NEJM* 345:1359,2001)
  - Activated protein C in severe sepsis (*NEJM* 344:699,2001)

# *Good Studies that Turn Out to Be Wrong*

---

[Ioannidis, *JAMA*, 294:218, 2005]

- Identified all 45 studies with positive results, published 1990-2003 in the “best” journals (highest impact factors), that had >1000 citations
- Searched for subsequent publication(s) on the same question with larger sample size or better study design
- Results -- 24% hadn't been redone, of those that were:
  - in 58% the original results *were* replicated
  - in 22% the subsequent findings contradicted the original study
  - in 22% the subsequent studies showed effect sizes  $\leq$  half that of the original study

## **WHY? $\Rightarrow\Rightarrow$ It's The Nature of Evidence**

---

- Consider the desire of the XYZ Corp to discover the effect of it's new, super-duper, gene-targeting antihypertensive agent, XYZ, on MAP in patients with hypertension
- The goal is to draw conclusions about the true effect of XYZ in the population of hypertensives  $\Rightarrow$  But we can't test it out in all people in the world with hypertension
- So we take a sample of hypertensive people, randomize them to get either XYZ or placebo, and try to make *inferences* about hypertensive people in general from this single sample
- This is why we need statistics -- if we could test out the drug vs. placebo in every person on Earth with hypertension, there'd be no need for statistical inference

## Question

---

- A RCT is performed comparing XYZ vs. placebo. The result shows that the mean MAP for XYZ is lower than that of placebo, with  $p=0.04$

*Which of the following is true?*

- A) There is a 96% chance that XYZ lowers MAP more than placebo
- B) There is a 4% chance that XYZ is no different from placebo in its effect on MAP
- C) Both
- D) Neither

# Sampling

---

- Consider the baseline MAP in the entire population and in our single, random sample of size  $N$  (called  $MAP_{pop}$ ,  $MAP_{sample}$ )
- If the sample size is large, we expect  $MAP_{pop}$  will be close to  $MAP_{sample}$ 
  - but even then, it's unlikely to be *exactly* the same
- If we made repeated random samples we'd expect a slightly different values each time
  - the statistical theory of sampling tell us that if we were to take samples of size  $N$  over and over, and plot their means, we will get a distribution of *sample means* that is a Gaussian curve whose mean is  $MAP_{pop}$  and whose spread is  $SD_{pop}/\sqrt{N}$

# Doing the Study

---

- When we do our study we :
  - take a single, random sample of hypertensive people  $\Rightarrow$
  - split them (randomly) into 2 subgroups  $\Rightarrow$
  - give half XYZ ( $\text{MAP}_{\text{sample,xyz}}$ ) and half placebo ( $\text{MAP}_{\text{sample,placebo}}$ )  $\Rightarrow$
  - do statistical analysis to address the question of whether XYZ is effective IN THE POPULATION by seeing its effect in the sample
  
- The standard approach to assessing the effect of XYZ is to consider a “null hypothesis” that in the *underlying population* there is no difference in MAP between XYZ and placebo (i.e.  $\text{MAP}_{\text{pop,xyz}} = \text{MAP}_{\text{pop,placebo}}$ )

# About the Null Hypothesis

---

- Even if the null hypothesis is true, then by the nature of random sampling it's unlikely that we'll find that  $MAP_{\text{sample,xyz}} = MAP_{\text{sample,placebo}}$
- Ask the question: *Assuming* that the null hypothesis is true (i.e.  $MAP_{\text{pop,xyz}} = MAP_{\text{pop,placebo}}$ ), what is the probability that in a sample of size N that we'd find a difference equal to the observed difference between  $MAP_{\text{sample,xyz}}$  &  $MAP_{\text{sample,placebo}}$  ?
- Say the answer is 4% (derived from the data)
- This **p-value** tells us that *IF* the null hypothesis were true (i.e. in the population), there was a 4% chance that in a single random sample of this size we would have found a difference between XYZ & placebo of what we actually found

# *What the p-value Is, & Isn't*

---

- So, the p-value says that if the null hypothesis is true, there was a 4% chance that we'd get what we actually found
- It does not tell us the probability that the null hypothesis is true (i.e. the probability that  $\text{MAP}_{\text{pop,xyz}} = \text{MAP}_{\text{pop,placebo}}$ ).
- Neither does it tell us the probability of whether the true effect of XYZ on MAP is different than that of placebo (e.g.  $\text{MAP}_{\text{pop,xyz}} \neq \text{MAP}_{\text{pop,placebo}}$ )
- In fact the p-value tells us **nothing at all** about the probability that there is truly a difference between the effect of XYZ and placebo on MAP on the underlying population

# *Inference Based on Null Hypothesis*

---

- **BUT WE MAKE A LEAP OF FAITH:** We have a *convention* (not a basic principle of truth), that if (under the null hypothesis) the probability of the observed results is small enough (i.e.  $p < 0.05$ ), then we choose to believe that the null hypothesis is wrong, and instead the truth is that  $MAP_{pop,xyz} \neq MAP_{pop,placebo}$
- BUT again, this conceptual construct does not actually tell us *anything at all* about the probability that:
  - $MAP_{pop,xyz} = MAP_{pop,placebo}$  or
  - $MAP_{pop,xyz} \neq MAP_{pop,placebo}$

## *The Limitations of this Leap of Faith*

---

- THUS -- we should not be surprised when a well designed, well executed, and well analyzed study that came up with a  $p < 0.05$  turn out to be untrue (i.e. are not confirmed by the weight of studies/evidence)
- Actually, the p-value is *fundamentally* incapable of indicating the strength of evidence for any hypothesis..... but.....
- There is a formalism that **CAN** do this: the Bayesian approach
  - most simply, the Likelihood Ratio (Bayes Factor) indicates the *relative* strength that a study provides for the null vs. alternate hypothesis
  - by adding to this a pre-study probability of what is true, we can use Bayes Theorem to derive the post-study probability of that truth

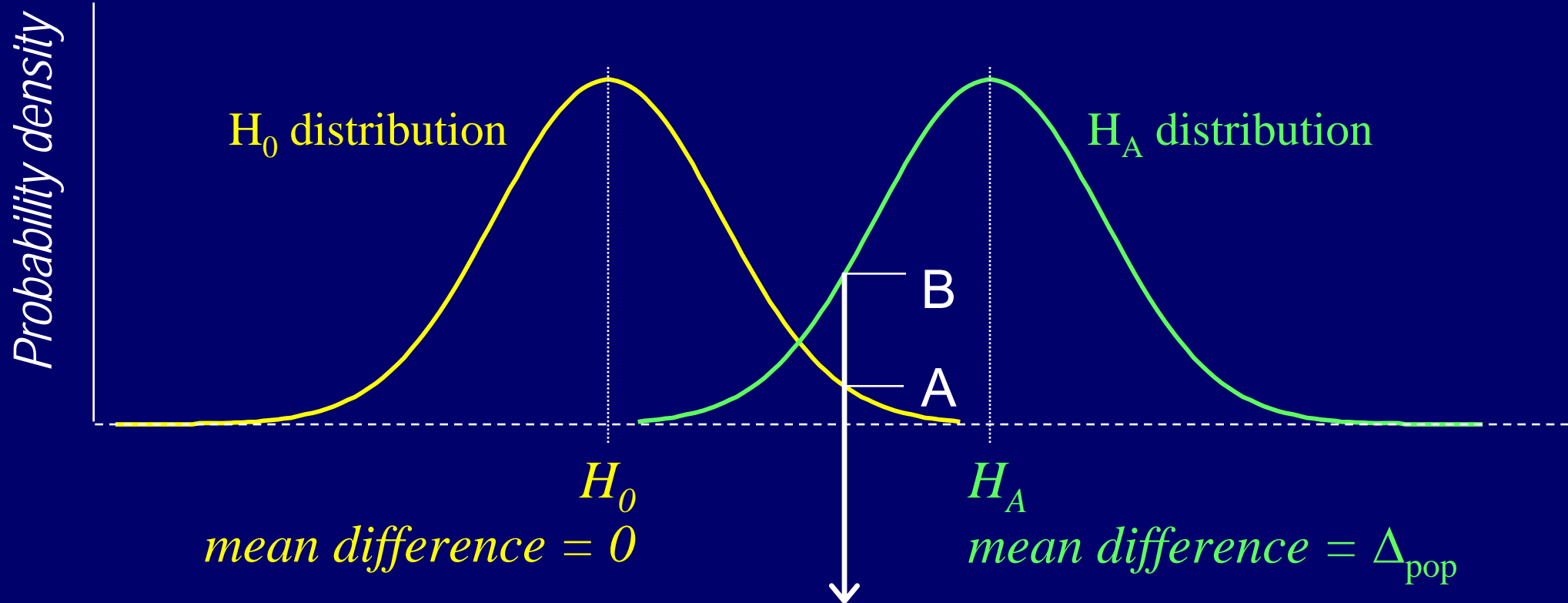
## *The Bayes Factor (LR)*

---

- Consider a continuous outcome, with a Gaussian distribution in the population
- Two hypotheses:
  - null hypothesis ( $H_0$ ) -- says the difference in the outcome between the 2 treatments in the population = 0
  - a *specific* alternative hypothesis ( $H_A$ ) -- says the difference in the population is a specific value =  $\Delta_{\text{pop}}$
- Do the study (in the sample)  $\Rightarrow$  find value of  $\Delta_{\text{sample}}$
- $LR = \text{Prob. of } \Delta_{\text{pop}} = \Delta_{\text{sample}} \text{ under } H_0 / \text{Prob of same under } H_A$

# The Bayes Factor (LR)

(Goodman *Am J Epidemiol* 137:491,1993)



Observed value of the difference =  $\Delta_{sample}$   
LR =  $A/B$  = experimental support for  $H_0:H_A$

(Note:  $A \neq \alpha$  and  $B \neq \beta$ )

# The Bayes Factor

- Obviously, every different alternative hypothesis gives a different value of LR
- Turns out that the **least** support for the null hypothesis occurs for the alternative hypothesis where  $\Delta_{\text{pop}} = \text{what was found } (\Delta_{\text{sample}})$ 
  - but even for *that* alternative hypothesis, the relative support for  $H_0$  is 3-5X smaller than the p-value, specifically:

<u>p-value</u>	<u>relative likelihood of <math>H_0</math> vs. <math>H_A</math></u>
0.05	0.15 → evidence for $H_A$ is ~7X (not 20X) that for $H_0$
0.03	0.10
0.01	0.04
0.001	0.005

- Thus, usual “significant” p-values provide much weaker evidence against the null hypothesis than most people realize

# *The Full Monte: Bayes Theorem*

---

- Null hypothesis ( $\alpha$ ) + Alternative hypothesis ( $\Delta$ ,  $\beta$ ) + Prestudy probability of what is true in the underlying population + Study results  $\rightarrow\rightarrow$  Poststudy probability of the truth
- It's exactly akin to clinical testing
- About the prestudy probability
  - many claim it's arbitrary and unscientific -- but it's no less arbitrary than the choice of  $p < 0.05$  to consider something to be true
  - in fact, it represents all our prior knowledge about the topic under study  $\Rightarrow$  puts the current study into it's most appropriate context
  - regarding hypothesis testing, Neyman & Pearson wrote: “*No test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis*”

# Summary & Conclusions

---

- Superior medical decision-making and medical care demands that we know, and use the best available medical evidence
- But there are many reasons that published studies can draw incorrect conclusions
- A substantial minority of studies are wrong even in the absence of deficiencies in study design, execution or analysis
- This is “because of the almost universal, but ill-founded practice of claiming truth based solely on finding a p-value  $< 0.05$ ” (or whatever threshold is used)
- We can do better (make less mistakes), using Bayesian analysis
  - and also more explicitly understand the nature and magnitude of our knowledge and uncertainty about what’s true

## References

---

- Goodman. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 137(5):491-501, 1993
- Goodman. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Ann Intern Med* 130(12):995-1004, 1999
- Goodman. Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. *Ann Intern Med* 130(12):1005-1013, 1999
- Ioannidis. Why Most Published Research Findings Are False. *PLoS Medicine* 2(8), 2005.
- Ioannidis. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *JAMA* 294(2):218-228, 2005