2026 ASCP Henry Nasrallah Award for Excellence in Clinical Psychopharmacology Education

The Tyranny of the P-value: Effect Size Matters

Leslie Citrome, MD, MPH

Clinical Professor of Psychiatry and Behavioral Sciences, New York Medical College

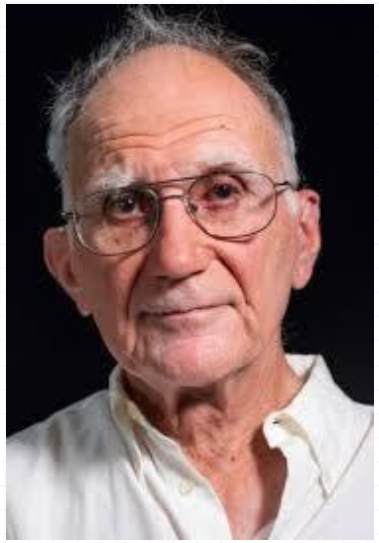
Faculty Disclosure

Leslie Citrome, MD, MPH, is a clinical professor in the Department of Psychiatry and Behavioral Sciences at New York Medical College in Valhalla, New York and adjunct professor at the Northwestern Feinberg School of Medicine in Chicago, Illinois.

Dr. Citrome has served as a consultant for Abbvie, Acadia, Adheretech, Altus, Alumis, Axsome, Alkermes, Arc, Auritec, Autobahn, Avant Healthcare Solutions, Biogen, BioXcel, BMS/Karuna, Boehringer Ingelheim, Cadent, Cerevel, Clario/Medavante/Prophase, Clinilabs, Compass, Corcept, Definium, Delpor, Eisai, Enteris BioPharma, Health Wellness Partners, HLS, Idorsia, Inmune Bio, Intra-Cellular, Johnson & Johnson/Janssen, Little Bear, Lundbeck, Luye, Lyndra, Maplight, Marvin, Mindmed, Neurelis, Neushen, Neumora, Neurocrine, Noema, Novartis, Noven, Orexo, Otsuka, Ovid, Pontifax/Draig, Praxis, PSL, Real Chemistry, Relmada, Renew Research, Response, Reviva, Sage, Seaport, Sumitomo/Sunovion, Supernus, Teva, University of Arizona, Vanda, Wells-Fargo, and one-off ad hoc consulting for individuals/entities conducting marketing, commercial, or scientific scoping research; speaker for Abbvie, Acadia, Alkermes, BMS, Eisai, Idorsia, Intra-Cellular, Johnson & Johnson/Janssen, Lundbeck, Luye, Neopharm, Neurocrine, Noven, Otsuka, Recordati, Takeda, Teva, Vanda, and CME activities organized by medical education companies such as Decera Clinical Education/Clinical Education Alliance/Clinical Care Options, CME Institute, CMEology, HMP/Psych Congress, Medscape/WebMD, MultiMedia Medical LLC, Neuroscience Education Institute, NEI, Paradigm, Real Psychiatry/Efficient, Real World CE, Rockpointe, Total CME, Vindico, and Universities, Professional Organizations/Societies and Advocacy Associations (MHA); owns small amounts of health-related shares of common stock in multiple companies in a portfolio managed externally, and stock options in Reviva; and earns royalties/publishing income from Taylor & Francis (Editor-in-Chief, Current Medical Research and Opinion, 2022-date), UpToDate (reviewer), Springer Healthcare (book), Elsevier (Topic Editor, Psychiatry, Clinical Therapeutics, through Spring 2025)

Acknowledgements

David McPherson



Jan Volavka

Jerome Levine

Chris Graf

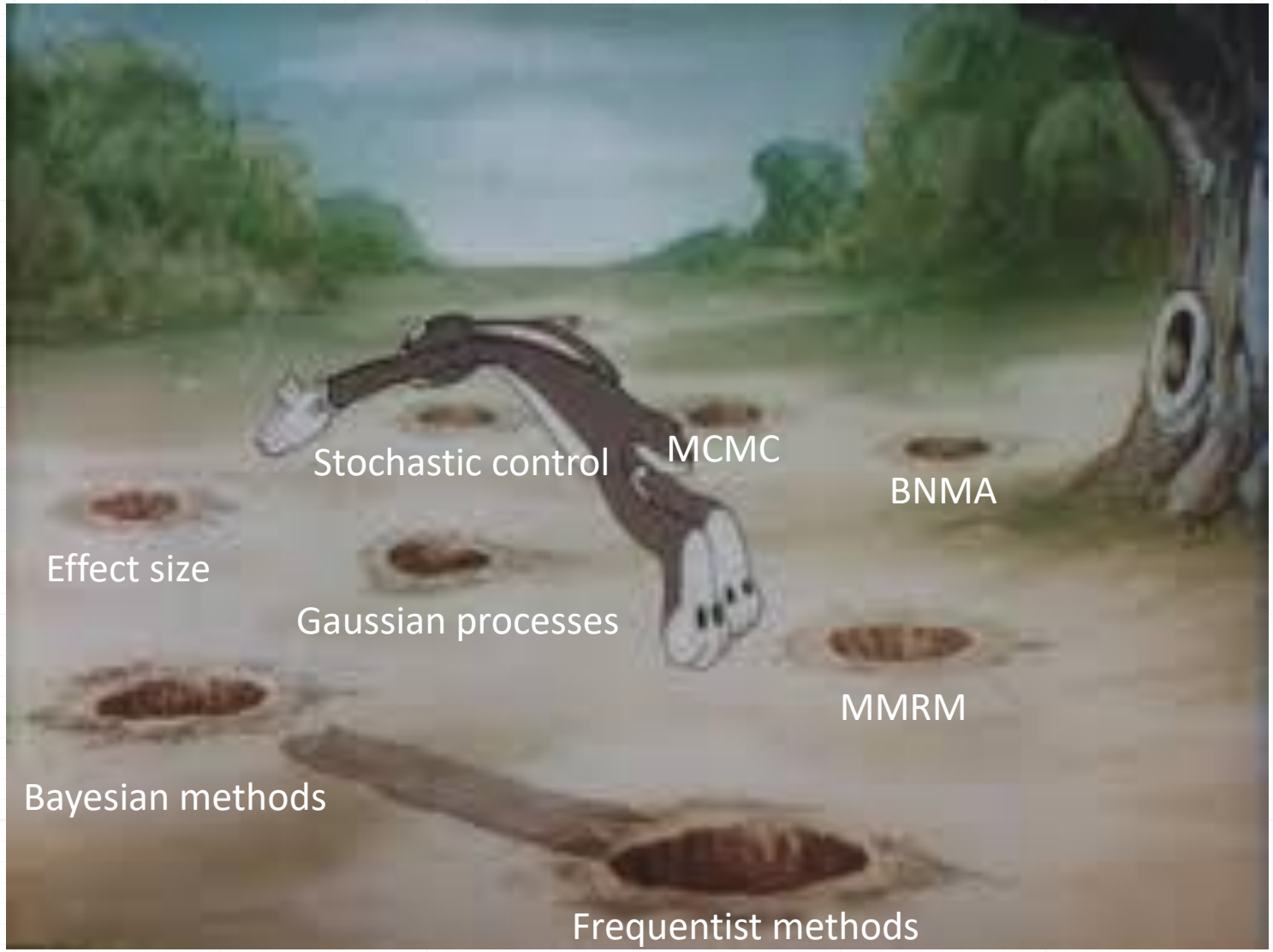
Jamie Karagianis

Stephen Stahl

Richard Davis

nanos gigantum humeris insidentes

Diving Into the Rabbit Holes of Statistics



“My Result Has a *P*-Value < .05”

Citrome L. Call for papers. *International Journal of Clinical Practice*. www.youtube.com/watch?v=ETM-8y68KbQ. Accessed June 7, 2019.

Why is This Presentation Being Done?

- Even smart people often assume that if a *P*-value < .05 then the result must be important
- Moreover, there is often the false belief that a *P*-value < .0001 denotes an even more impressive difference
- There is a general lack of understanding of effect size and how this can be used to appraise clinical trial results
- Simple to calculate measures of effect size such as number needed to treat (NNT) and number needed to harm (NNH) can help treaters place study outcomes in a more relevant clinical context

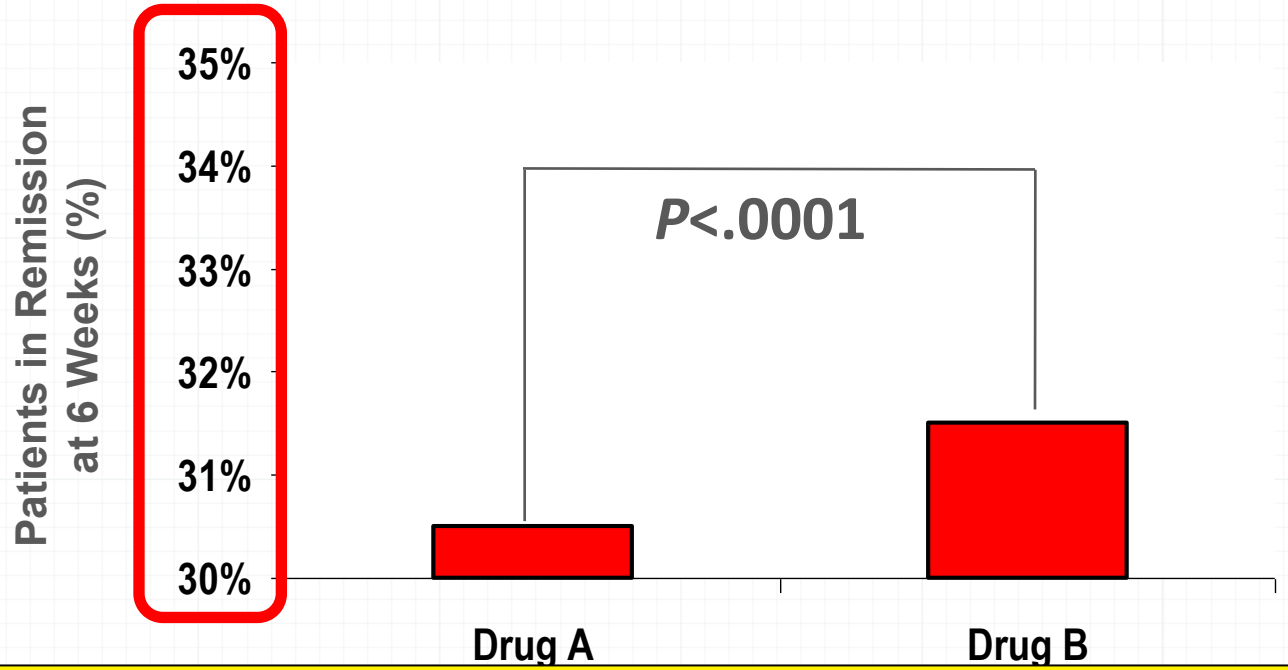
Outline

- What is an effect size and how is it different from a *P*-value?
- A word about relative effect sizes
- What are Number Needed to Treat (NNT) and Number Needed to Harm (NNH)?
- What is Likelihood to be Helped or Harmed (LHH)?
- Summary

Outline

- What is an effect size and how is it different from a *P*-value?
- A word about relative effect sizes
- What are Number Needed to Treat (NNT) and Number Needed to Harm (NNH)?
- What is Likelihood to be Helped or Harmed (LHH)?
- Summary

The difference in remission for a major depressive episode at 6 weeks for Drug A vs Drug B is highly statistically significant, but clinically irrelevant



How irrelevant is this? Can we quantify this?

Citrome L. *Acta Psychiatr Scand*. 2008;117(6):412-419.

Concepts Related to Benefit / Risk:
P-Value

- This gives an indication of how strong the likelihood that any difference is NOT due to chance
- The smaller the *P*-value, the more convinced you are that something is going on that is not just random
- This does not state anything about the size or the importance of the non-random effect
- *P*-value is not the same as effect size

Citrome L. *Acta Psychiatr Scand*. 2010;121(2):94-102.

Concepts Related to Benefit / Risk:
Effect Size

- This gives an indication of how big the treatment effect is in terms of reduction in symptoms, or other outcome of interest
- The greater the effect size, the more convinced you are that the intervention will have a clinically important impact
- This does not state anything about statistical significance of the observed outcome in a clinical trial
- Effect size is not the same as *P*-value

Citrome L. *Acta Psychiatr Scand*. 2010;121(2):94-102.

There are Several Measures of Effect Size

Effect Size	Range of Possible Values (weakest, ie, no difference, to strongest)	Typical Example of a Small Effect	Typical Example of a Large Effect
Relative Measures			
Relative Risk	1 to ∞	2	4
Odds Ratio	1 to ∞	2	4
Hazard Ratio	1 to ∞	2	4
Relative Risk Increase	1 to ∞	< 100%	300%
Absolute Measures			
Attributable Risk	0 to 100%	< 10%	33%–50%
NNT	∞ to 1	≥ 10	2–3
Cohen's *d*	0 to ∞	0.2	0.8
Area Under the Curve	0.50 to 1.00, or 0.50 to 0	0.56	0.71
Success Rate Difference	0 to 1	0.11	0.43

Citrome L. *Acta Psychiatr Scand*. 2010;121(2):94-102.

P-Value vs Effect Size

P-Value	Effect Size
Indicates Statistical Significance	Indicates Clinical Significance
Independent of Effect Size	Independent of *P*-Value

Outline

- What is an effect size and how is it different from a *P*-value?
- A word about relative effect sizes
- What are Number Needed to Treat (NNT) and Number Needed to Harm (NNH)?
- What is Likelihood to be Helped or Harmed (LHH)?
- Summary

Contrasting Absolute and Relative Risk

Prospective Results from the Women's Health Initiative

BBC NEWS

Aspirin cuts breast cancer risk

A new piece of US research backs the idea that aspirin protects against certain types of breast cancer.

It found women who used aspirin or similar painkillers at least once per week for six months reduced their risk of breast cancer by 20%.

http://news.bbc.co.uk/go/pr/fr/-/2/hi/health/3748697.stm

Contrasting Absolute and Relative Risk

Prospective Results from the Women's Health Ini

BBC NEWS

Aspirin cuts br

A [illegible] idea that aspirin protects against certain types of breast

[illegible]who used aspirin or similar painkillers at least once per week for six months reduced their risk [illegible] cancer by 20%.

Statistically significant, but clinically irrelevant!
Very small effect size!

http://news.bbc.co.uk/go/pr/fr/-/2/hi/health/3748697.stm

Contrasting Absolute and Relative Risk

Prospective Results from the Women's Health Initiative

Healthy Skepticism | *Part of an occasional series*

Overstating Aspirin's Role In Breast Cancer Prevention

How Medical Research Was Misinterpreted to Suggest Scientists Know More Than They Do

By Lisa M. Schwartz, Steven Woloshin and H. Gilbert Welch
Special to The Washington Post

Medical research often becomes news. But sometimes the news is made to appear more definitive and dramatic than the research warrants. This series dissects health news to highlight some common study interpretation problems we see as physician researchers and show how the research community, medical journals and the media can do better.

Preventing breast cancer is arguably one of the most important priorities for women's health. So when the Journal of the American Medical Association published research a year ago suggesting that aspirin might lower breast cancer risk, it was understandably big news. The story received extensive coverage in top U.S. newspapers, including The Washington Post, the Wall Street Journal, the New York Times and USA Today, and the major television networks. The headlines were compelling: "Aspirin May Avert Breast Cancer" (The Post), "Aspirin Is Seen as Preventing Breast Tumors" (the Times).

In each story, the media highlighted the change in risk associated with aspirin — noting prominently something to the effect that aspirin users had a "20 percent lower risk" compared with nonusers. The implied message in many of the stories was that women should consider taking aspirin to avoid breast cancer.

But the media message probably misled readers about both the size and certainty of the benefit of aspirin in preventing breast cancer. That's because the reporting left key questions unanswered:

See ASPIRIN, *Page F4*

"In each story, the media highlighted the change in risk associated with aspirin -- noting prominently something to the effect that aspirin users had a "20 percent lower risk" compared with nonusers."

"The implied message in many of the stories was that women should consider taking aspirin to avoid breast cancer."

Schwartz LM et al. The Washington Post Tuesday, May 10, 2005.

Contrasting Absolute and Relative Risk

Prospective Results from the Women's Health Initiative

- Absolute risk
  - The risk of developing breast cancer for postmenopausal women who do not take aspirin on a regular basis is 955/194, 884 person-years, or 0.49%
- Relative risk
  - Taking an aspirin a day for at least 5 years reduces risk by 20% to 99/24,398 person-years, or 0.41%; this is a relative risk reduction of 20%
- The absolute risk reduction is only 0.08% versus a relative risk reduction of 20%

Harris RE et al. Cancer Research 2003;63:6096-6101.

Contrasting Absolute and Relative Risk

Prospective Results from the Women's Health Initiative

- "Another way to present these results would be to say that a woman's chance of being free from breast cancer over the next five years was 98.4 percent if she used aspirin and 98 percent if she did not.
- "Seeing the actual risks leaves a very different impression than a statement like 'aspirin lowers breast cancer risk by 20 percent.' "

Schwartz LM et al. The Washington Post Tuesday, May 10, 2005.

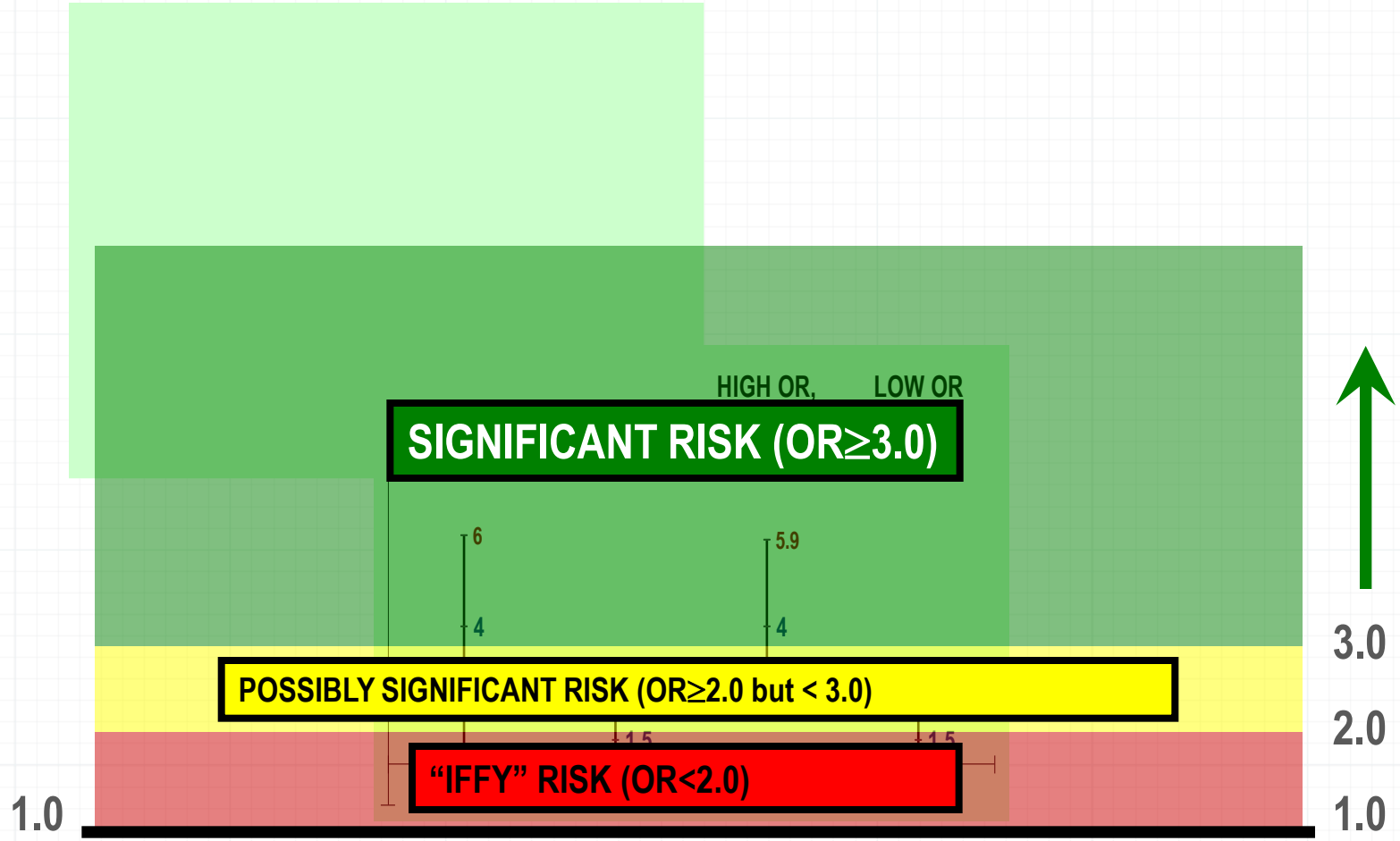
Relative Measures: Rules of Thumb

- In observational studies, many epidemiologists will be less likely to accept at face value an RR of less than 3.0, and almost never an RR of less than 2.0
- Increases in risk of less than 100% (i.e. an RR of less than 2.0) are difficult to interpret because of the possibility of confounding factors that have not been accounted for

Citrome L. Driugs of Today 2004;40(5):445-464.

Relatively Speaking, How Odd Can Hazards Be?

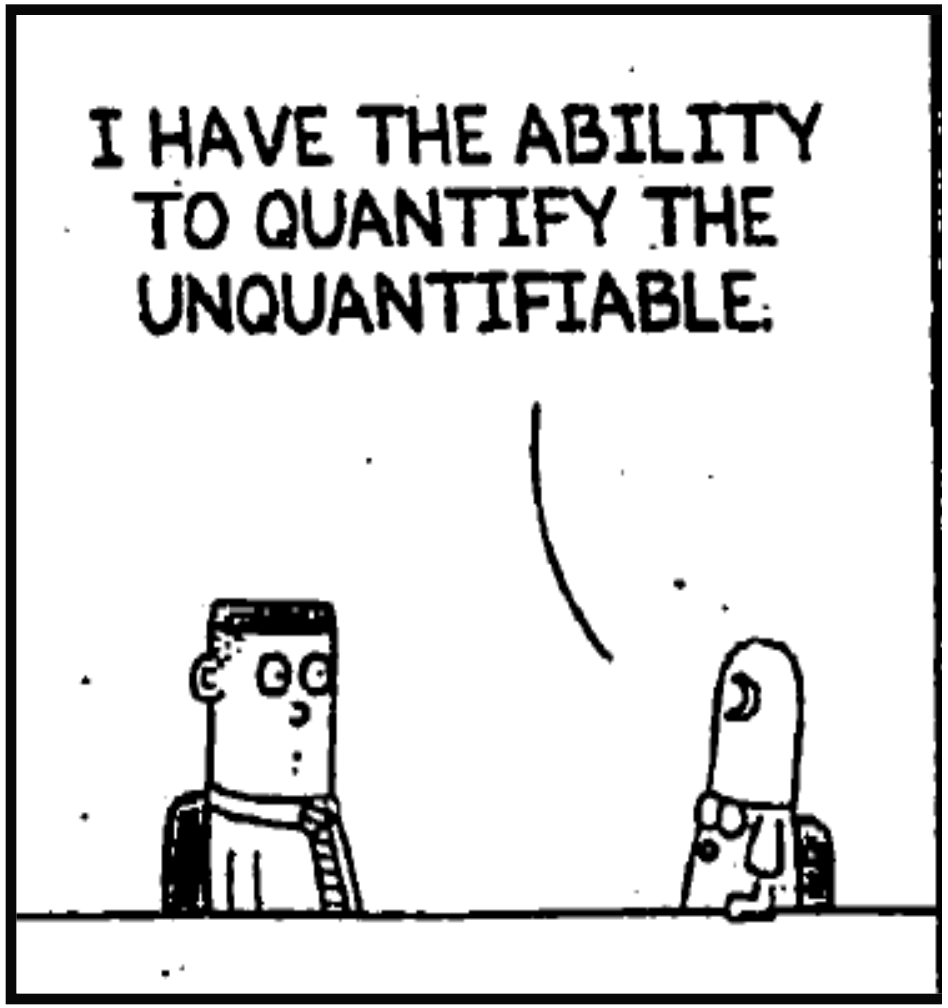
Strength of the association may be high, but statistical significance may not be reached (if confidence intervals includes 1.0)



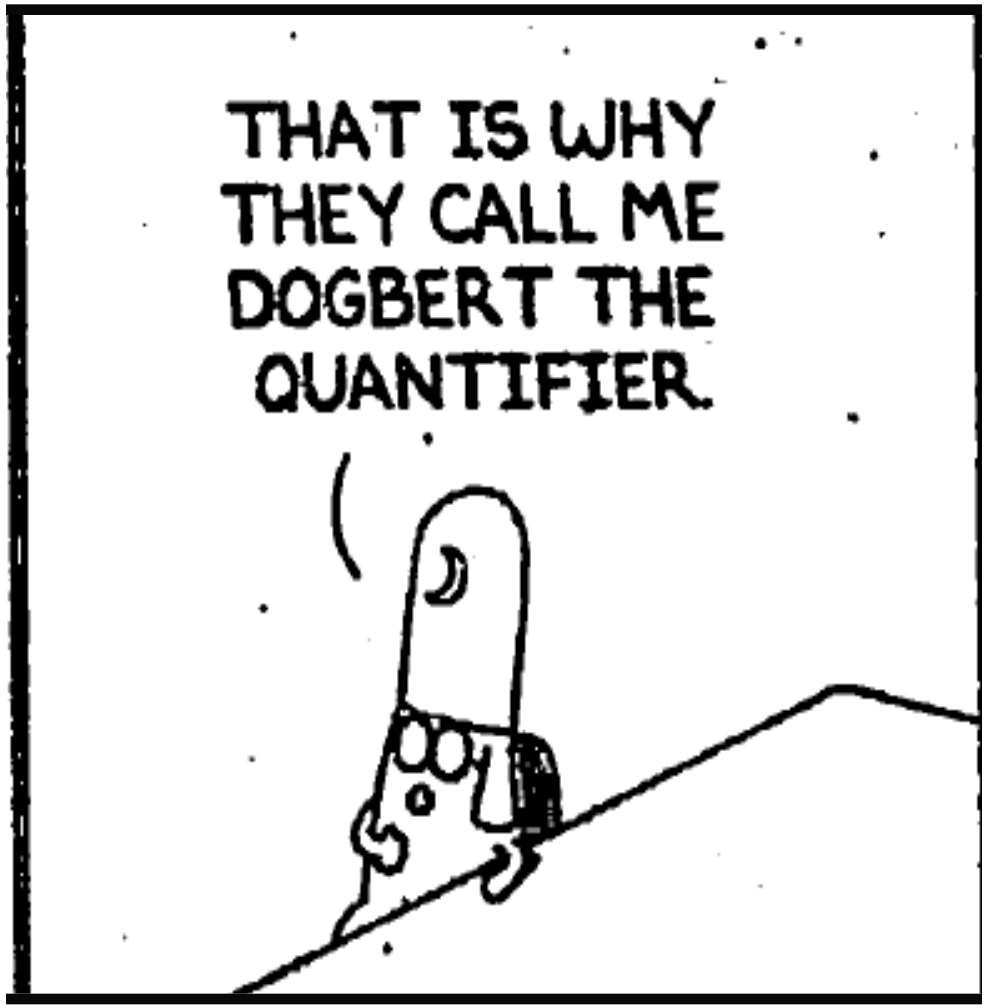
Citrome L. PsychEdUp 2005;1(10):2-3.

Outline

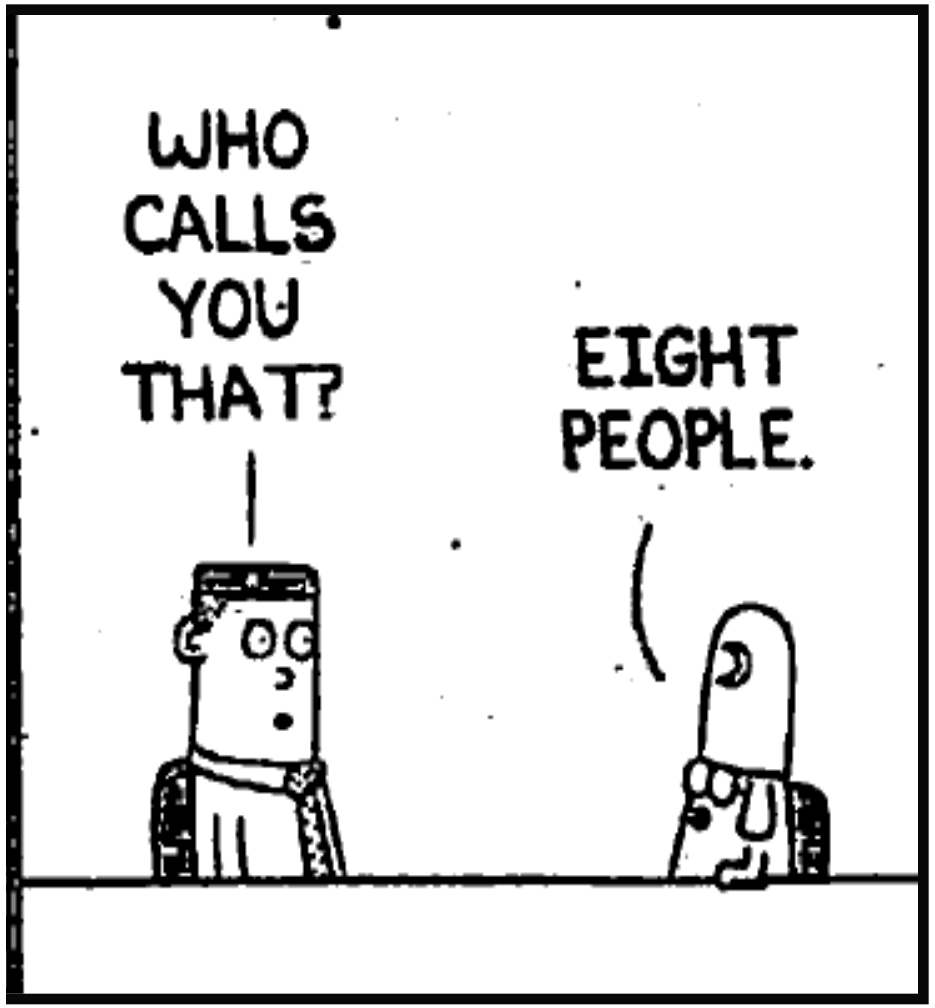
- What is an effect size and how is it different from a *P*-value?
- A word about relative effect sizes
- What are Number Needed to Treat (NNT) and Number Needed to Harm (NNH)?
- What is Likelihood to be Helped or Harmed (LHH)?
- Summary



Daily News April 5, 2007

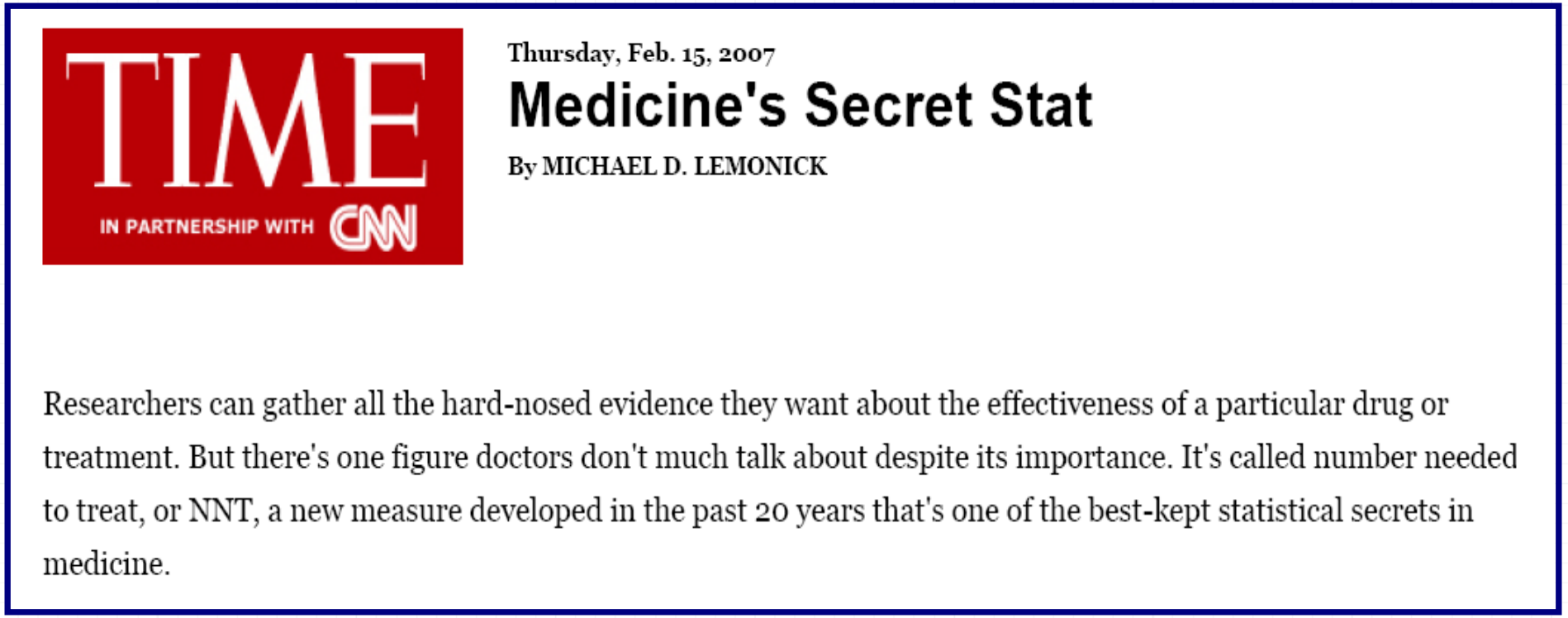


Daily News April 5, 2007



Daily News April 5, 2007

Medicine’s Impact on Society



TIME
IN PARTNERSHIP WITH CNN

Thursday, Feb. 15, 2007

Medicine's Secret Stat

By MICHAEL D. LEMONICK

Researchers can gather all the hard-nosed evidence they want about the effectiveness of a particular drug or treatment. But there's one figure doctors don't much talk about despite its importance. It's called number needed to treat, or NNT, a new measure developed in the past 20 years that's one of the best-kept statistical secrets in medicine.

http://www.time.com/time/magazine/article/0,9171,1590464,00.html

Concepts Related to Benefit / Risk:
Effect Size – NNT

- NNT is one measure of effect size
- It is independent of *P*-value and does not say anything about the likelihood of the difference between treatments being due to chance alone
- Helps you judge the clinical significance of a statistically significant result

Citrome L. *Acta Psychiatr Scand*. 2010;121(2):94-102.

NNT

- How many patients would you need to treat with Drug A instead of Drug B before you would encounter 1 extra outcome of interest, such as response

Small NNT Numbers = Bigger Difference between Drug A and Drug B

Citrome L. *Acta Psychiatr Scand*. 2010;121(2):94-102.

Calculating NNT is Easy

What is the NNT for an outcome for Drug A vs Drug B?

f_A = frequency of outcome for Drug A
f_B = frequency of outcome for Drug B
Attributable Risk (AR) = $f_A - f_B$
NNT= 1/AR

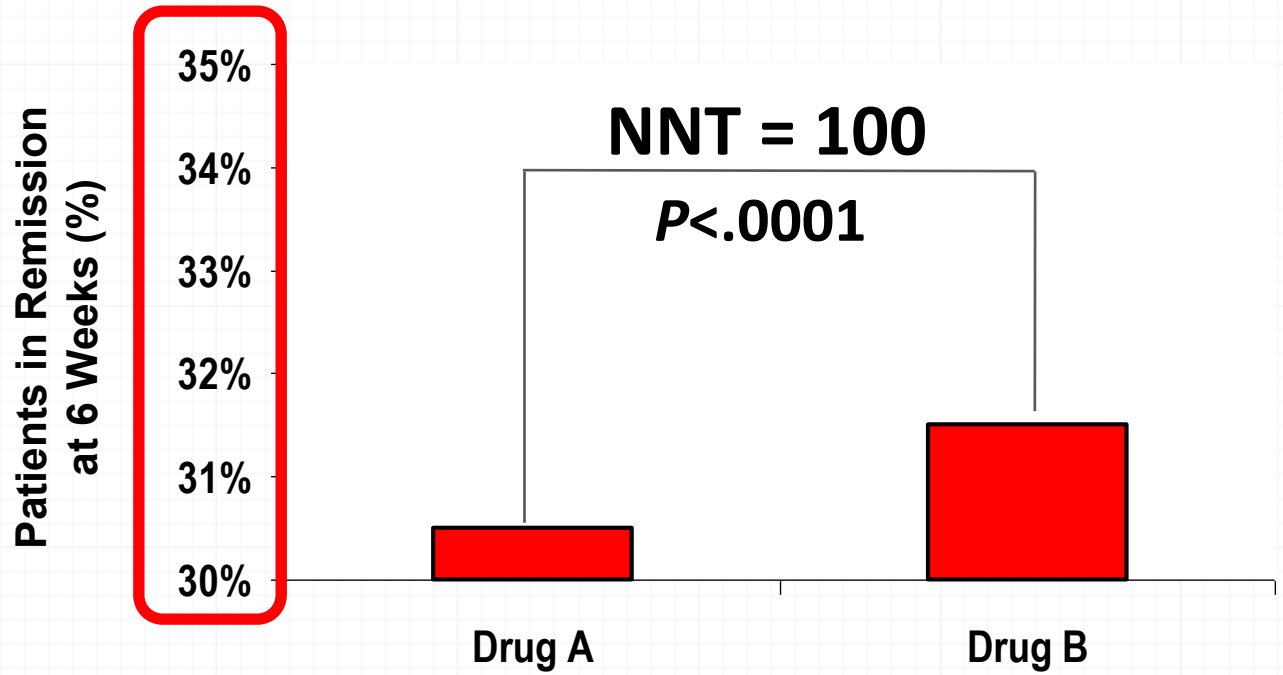
By convention, when not presenting fractions, we round up the NNT to the next *higher* whole number

For example, Drug A results in remission 50% of the time, but Drug B results in remission 20% of the time.

NNT = 1/[0.50-0.20] = 1/0.30 = 3.33 → Round up to 4

Citrome L. *Acta Psychiatr Scand*. 2010;121(2):94-102.

Back to Our Example: Quantifying Irrelevance



NNT = 1/(0.315-0.305) = 1/0.01 = 100

Citrome L. *Acta Psychiatr Scand*. 2008;117(6):412-419.

What is NNH?

- NNH is Number Needed to Harm
- NNH is calculated the same way as NNT
- We would use the term NNH when referring to an outcome we are trying to avoid, or to refer to a disadvantage for Drug A vs Drug B
- In calculating NNT, if it is a negative number, we can call it an NNH

Citrome L. *Acta Psychiatr Scand*. 2008;117(6):412-419.

NNT

NNH

What is a Clinically Important NNT?

- A large NNT of ≥ 100 means that there is little difference between choosing Drug A or Drug B for the outcome measured
- A small NNT of 2 would be a hugely important difference
- NNT values < 10 denote a potentially useful intervention
- NNH values > 10 denote a potentially tolerable intervention
- Some NNTs may be clinically important, even if they are relatively large, for example when the outcome is death

Citrome L. *Acta Psychiatr Scand.* 2008;117(6):412-419.

Examples of NNT from Cochrane Reviews

NNT	Treatment, Disorder
2	Antibiotics versus placebo for preventing recurrent urinary tract infection in non-pregnant women (Albert et al., 2004)
3	Colchicine versus placebo to reduce pain in acute gout (Schlesinger, Schumacher, Catton, & Maxwell, 2006)
4	Anticonvulsants versus placebo for migraine prophylaxis (Chronicle & Mulleners, 2004)
5	Antibiotics versus placebo for cough relief in acute bronchitis (Smucny, Fahey, Becker, & Glazier, 2004)
6	Fluocinonide versus hydrocortisone cream for discoid lupus erythematosus (Jessop, Whitelaw, & Jordaan, 2001)
7	Antibiotics (anti-H. pylori) versus antisecretory therapy for prevention of recurrent bleeding from peptic ulcer (Gisbert et al., 2004)
8	Single dose dextropropoxyphene versus placebo for postoperative pain (Collins, Edwards, Moore, & McQuay, 2000)
9	Antifungals versus placebo for preventing oral candidiasis in patients with cancer receiving treatment (Clarkson, Worthington, & Eden, 2007)
10	Decongestant plus antihistamine for acute otitis media in children (Flynn, Griffin, & Schultz, 2004)
15	Carotid endarterectomy versus not for symptomatic carotid stenosis (Cina, Clase, & Haynes, 2000)
20	Oxygen versus room air resuscitation for preventing death in infants at birth (Tan, Schulze, O'Donnell, & Davis, 2005)

Can We Express Statistical and Clinical Significance Together?

- We can do this for NNT by also giving the CI
  - What is the range of values of NNT within which “the truth” probably exists?
  - If this range includes “infinity” it means it can take an infinite number of patients to see a difference, ie, there is no difference
  - CI tells us about the precision of our estimate of NNT
- You can calculate it with a simple formula, or use an online calculator

Citrome L. *Acta Psychiatr Scand*. 2008;117(6):412-419.

Limitations of Using NNT / NNH

- GIGO
- It is most valid to calculate from an RCT with identical conditions for all drugs under study
- Results are only calculable for binary or dichotomous events that are either present or absent, and do not apply to continuous variables such as the value of a blood test
- However, values with clinically significant thresholds, such as weight gain > 7% can be expressed as an NNT because then they are binary

Citrome L. *Acta Psychiatr Scand*. 2008;117(6):412-419.

NNT Summary

- Absolute differences place the data in a clinically meaningful context, relative differences can be deceiving
- The concept of NNT allows the clinician to estimate a medication's potential relevant effect
- Examining the magnitudes of NNT (and NNH), the clinician can start to make risk–benefit decisions tailored to the individual patient's needs or preferences

Citrome L. *Acta Psychiatr Scand*. 2008;117(6):412-419.

Outline

- What is an effect size and how is it different from a *P*-value?
- A word about relative effect sizes
- What are Number Needed to Treat (NNT) and Number Needed to Harm (NNH)?
- What is Likelihood to be Helped or Harmed (LHH)?
- Summary

LHH:
Likelihood to be Helped or Harmed

- LHH ratios (likelihood of being helped or harmed) can be a valid and useful way of synthesizing data regarding benefits and risks
- LHH = (1/NNT) / (1/NNH) = NNH/NNT
- An LHH > 1 would mean the likelihood to be helped is greater than the likelihood to be harmed; for an LHH < 1, the reverse is true
- Benefit–risk can thus be quantified by calculating LHH, provided that the benefit and harm being considered are relevant to the particular individual and are logically matched in terms of expected time course and consequences
- Remember that benefits and risks can take on greatly differing degrees of importance or relevance depending on the subjective point of view of the patient and clinician, baseline risks, and severity of the underlying illness

Citrome L, et al. *Exp Rev Neurother*. 2008;8(7):1079-1091. Straus SE. *Eval Health Prof*. 2002;25(2):210-224.

Deutetrabenazine vs Valbenazine: NNT in Fixed-Dose Trials - Very Similar

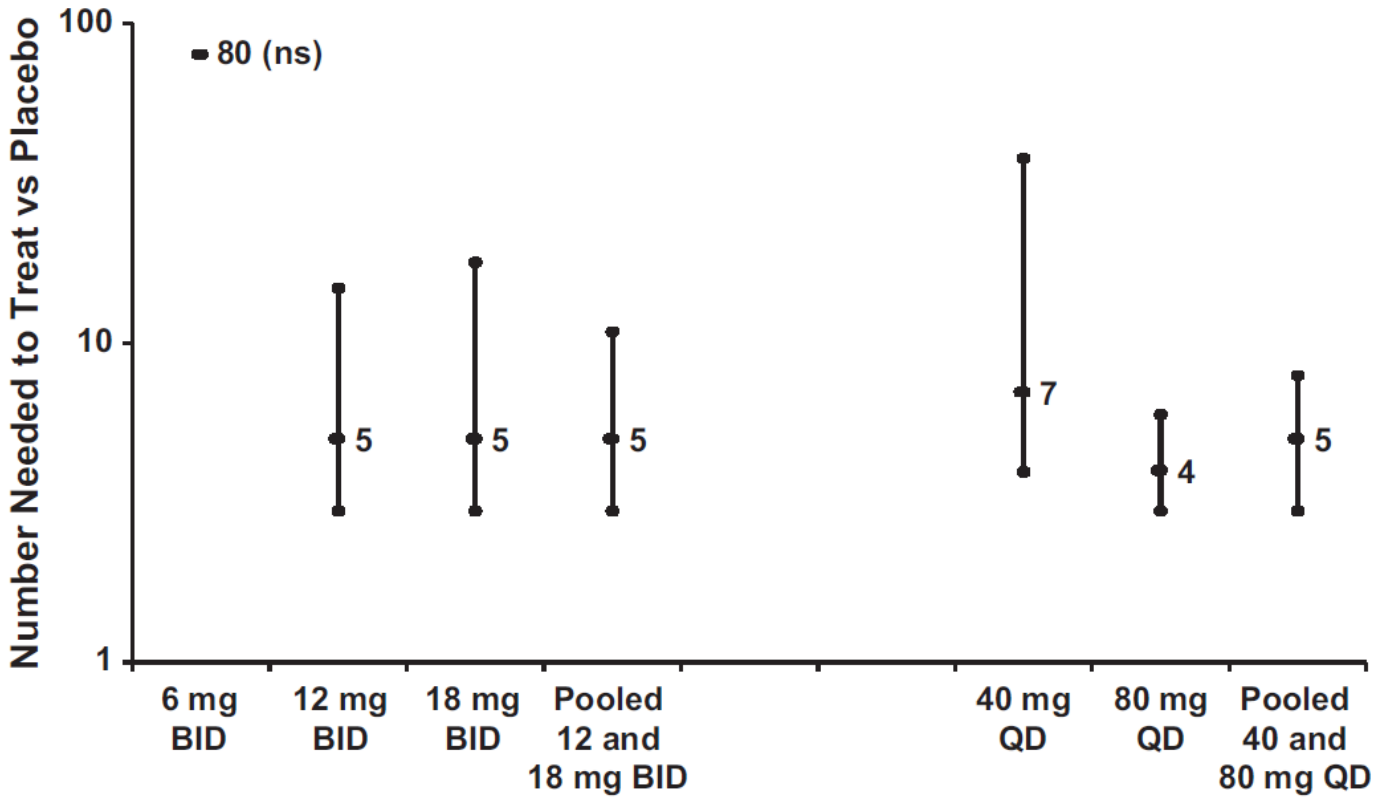
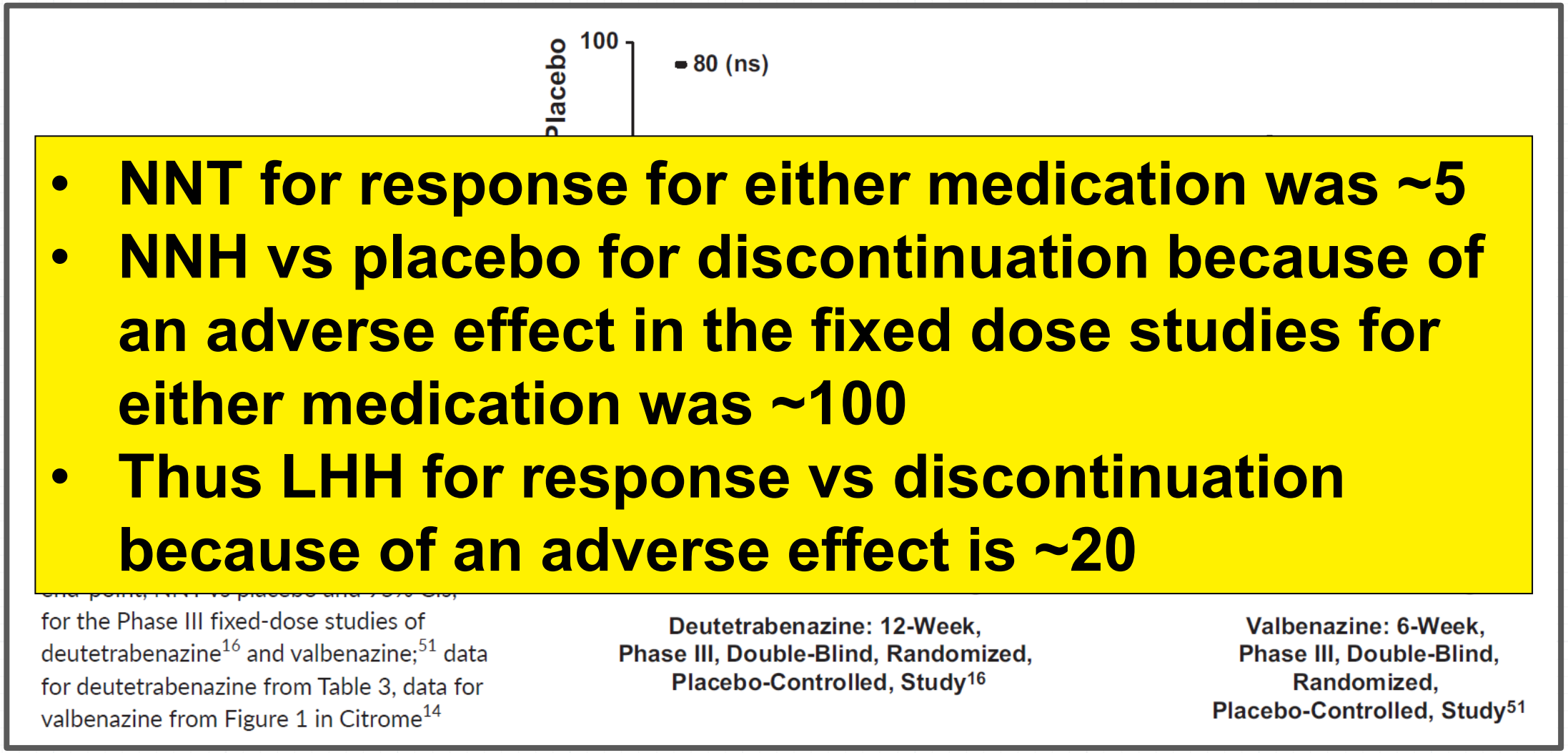


FIGURE 1 Reduction (≥50%) in AIMS dyskinesia score from baseline to end-point, NNT vs placebo and 95% CIs, for the Phase III fixed-dose studies of deutetrabenazine[16] and valbenazine;[51] data for deutetrabenazine from Table 3, data for valbenazine from Figure 1 in Citrome[14]

Citrome L, et al. *Int J Clin Pract.* 2017;71:e13030. Citrome L. *J Neurol Sci.* 2017;383:199-204.

Deutetrabenazine vs Valbenazine: NNT in Fixed-Dose Trials – Very Similar



- **NNT for response for either medication was ~5**
- **NNH vs placebo for discontinuation because of an adverse effect in the fixed dose studies for either medication was ~100**
- **Thus LHH for response vs discontinuation because of an adverse effect is ~20**

for the Phase III fixed-dose studies of deutetrabenazine[16] and valbenazine;[51] data for deutetrabenazine from Table 3, data for valbenazine from Figure 1 in Citrome[14]

Deutetrabenazine: 12-Week, Phase III, Double-Blind, Randomized, Placebo-Controlled, Study[16]

Valbenazine: 6-Week, Phase III, Double-Blind, Randomized, Placebo-Controlled, Study[51]

Citrome L, et al. *Int J Clin Pract.* 2017;71:e13030. Citrome L. *J Neurol Sci.* 2017;383:199-204.

Outline

- What is an effect size and how is it different from a *P*-value?
- A word about relative effect sizes
- What are Number Needed to Treat (NNT) and Number Needed to Harm (NNH)?
- What is Likelihood to be Helped or Harmed (LHH)?
- Summary

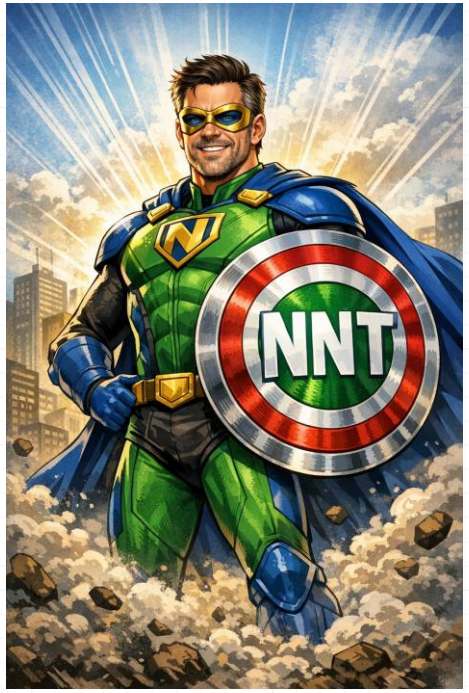
Summary

- NNT and NNH can be easily calculated
- NNT and NNH must be interpreted in a clinical context
  - Every patient has individual patterns of treatment and a different set of preferences
- LHH can help when discussing trade-offs between benefit and risk, provided both are meaningful to the clinician and patient, and are comparable in terms of time course

Human artist rendition 1996

Questions?

nntman@gmail.com



AI rendition 2026