How to Read a Journal Article*

James V Nepola, M.D.
Professor of Orthopaedics and Rehabilitation
Carver College of Medicine
University of Iowa
Iowa City, IA

*Thanks to Dr. Chad A. Krueger M.D for his assistance

There is Much to Read

- 4177 peer reviewed articles (2019)
- 10 Major Orthopaedic Journals
  - JBJS (Am)
  - CORR
  - AJSM, JSES, J Arthroplasty, Spine, JOR, JOT, Hand, JPOS
- ALMOST UNREADABLE
- NOT ALL “MUST READS”

What to Read “a strategy”

- Identify a “trusted” general orthopaedic journal (i.e. JBJS)
- 1 or 2 Specialty journals of interest
- Set aside a “reading time goal” each week
- Scan the journals
- Decide what is “relevant” to your situation/practice
- Scan abstracts and ask the question, “If the conclusions were true would it matter to my practice and/or patients?”

“Conversation Starters”

“There should be room for [reading] what we like to call “conversation-starters,” which we define as papers that open our minds to new ways of thinking about education or practice, but that don’t expose patients to serious risk.”

S. S. Leopold, Editor-in-Chief, *Clinical Orthopaedics and Related Research®*

Orthopaedic Research—We Know the Problem. What is the Solution? CORR Sept 2018 476(9);1689-91

How do I know if the article is worthwhile?

P.I.C.O.T. process for article assessment

P. – Patient population or problem of interest

I. – Investigational intervention or exposure

C. – Comparison Group, Control, Placebo, standard of care

O. – Outcome of Interest

T. – Study time, Appropriate follow-up

“Article Triaging”?

- “External Validity” – The P. and the I.
- Is it relevant to my needs or those of my patient population?
- Is the question asked concise and practical – does it answer my clinical problem
- If the answer is “no”, move on

“Article Triaging”

- “Internal Validity”
- Read the “Methods” The C . - Comparison
- What type of article is it?
    - Randomized
    - Controlled
    - Retrospective
    - Prospective
- “Are the results meaningful” The O - Outcome
- Does the follow up and timing make sense – The T

Bhandari et al., User’s guide to the orthopaedic literature: how to use an article about surgical therapy. JBJSAM. 2001 Jun: 83(6);916-26

“Article Triaging” - Bias

- Review the Author’s potential “Conflicts of Interests”
- If familiar with the topic, review whether the author or the author’s institution is an advocate of the technique or treatment being studied
  - This may inject a bias either in attitude regarding or in familiarity with a particular technique biasing the results and conclusions
  - For example: An institution which advocates a treatment and performs hundreds of a given procedure a year may report more positive results than would be experienced by the reader

“Article Triaging” - Reader Bias

- **Confirmation Bias** - tendency to interpret new evidence as confirmation [or refutation] of one's existing beliefs or theories.
  - Reader driven
  - Article affirms their bias born of training or habit - accept
  - Article contradicts their point of view - reject
  - Readers must keep an open mind and carefully and objectively evaluate the manuscript while realizing their own bias and doing their best to control it

Levels of Evidence (LoE)- AAOS 2003

1. Randomized Control Study or Meta-analysis of Level 1 “Homogeneous” RCT’s
2. Prospective Cohort Study or Poor Quality RCT with <80% FU
3. Case Control Study

   Retrospective Cohort Study
4. Case Series without Controls
5. “Expert Opinion”

Levels of Evidence (LoE)- AAOS 2005
Based on Primary Research Question

Therapeutic Studies: Investigating the results of treatment

Level I	- High quality randomized clinical trial (RCT) with statistically significant difference or no statistically significant difference but narrow confidence intervals - Systematic review of Level I RCTs with homogenous results
Level II	- Lesser quality RCT (e.g. < 80% follow-up, no blinding, or improper randomization) - Prospective comparative study - Systematic review of Level II studies or Level 1 studies with inconsistent results
Level III	- Case control study - Retrospective comparative study - Systematic review of Level III studies
Level IV	- Case series
Level V	- Expert opinion

Modified from Levels of Evidence and Grades of Recommendations by James G Wright
http://www2.aaos.org/bulletin/apr05/fline9.asp

Levels of Evidence (LoE)- AAOS 2005
Based on Primary Research Question

Prognostic Studies: Effect of patient characteristic on outcomes

Level I	- High quality prospective study (all patients were enrolled at the same point in their disease with ≥ 80% follow-up of enrolled patients) - Systematic review of Level I studies
Level II	- Retrospective study - Untreated controls from an RCT - Lesser quality prospective study (e.g. patients enrolled at different points in their disease or <80% follow-up.) - Systematic review of Level II studies
Level III	- Case control study
Level IV	- Case series
Level V	- Expert opinion

Modified from Levels of Evidence and Grades of Recommendations by James G Wright
http://www2.aaos.org/bulletin/apr05/fline9.asp

Levels of Evidence (LoE)- AAOS 2005
Based on Primary Research Question

Diagnostic Studies: Investigating a diagnostic test

Level I	- Testing of previously developed diagnostic criteria on consecutive patients (with universally applied reference "gold" standard) - Systematic review of Level I study
Level II	- Development of diagnostic criteria on consecutive patients (with universally applied reference "gold" standard) - Systematic review of Level II studies
Level III	-Study of non-consecutive patients; without consistently applied reference "gold" standard - Systematic review of Level III studies
Level IV	- Case control study - Poor reference standard
Level V	- Expert opinion

Modified from Levels of Evidence and Grades of Recommendations by James G Wright
http://www2.aaos.org/bulletin/apr05/fline9.asp

Levels of Evidence (LoE)- AAOS 2005
Based on Primary Research Question

Economic and Decision Analyses: Developing an economic or decision mode

Level I	- Sensible costs and alternatives; values obtained from many studies; with multiway sensitivity analyses - Systematic review of Level I studies
Level II	- Sensible costs and alternatives; values obtained from limited studies; with multiway sensitivity analyses - Systematic review of Level II studies
Level III	- Analyses based on limited alternatives and costs; and poor estimates - Systematic review[2] of Level III studies
Level IV	- Analyses with no sensitivity analyses
Level V	- Expert opinion

Modified from Levels of Evidence and Grades of Recommendations by James G Wright
http://www2.aaos.org/bulletin/apr05/fline9.asp

What is the Level of Evidence?

- A prospective study comparing PRP and cancellous allograft to autologous bone graft for treatment of nonunions of the proximal tibia stabilized with locking plates.
- All were closed fractures without infection. All patients were treated with same protocol other than type graft material at the same hospital. There was no randomization .
- To reduce selection bias the two treatment cases were matched appropriately and >80% of patients were followed for 18 months.
- The study concluded that there was no difference in the percentage of healing using standardized radiographic analysis at 6 months.

-

What is the level of evidence of the study?

Level 1 2 3 4 5

What is the level of evidence?

Level 2 - A Prospective Cohort Therapeutic Study with > 80% follow up

Level - II therapeutic studies include all the following:

1. well-conceived prospective cohort studies,
2. poor-quality randomized controlled trials (i.e. follow-up less than 80%)
3. systematic reviews of Level-II studies or non-homogenous Level-I studies.

Percentage of Level 1 Evidence Published Trauma Manuscripts (JOT, JBJS, CORR) between 2013 to 2018

- Journal of Orthopaedic Trauma – 5% to 2%
- Clinical Orthopaedics and Related Research – 5% to 15%
- Journal of Bone and Joint Surgery:AM – 20% to 18%

Luksameearunothai, K., Chaudhry, Y., Thamyongkit, S. *et al.* Assessing the level of evidence in the orthopaedic literature, 2013–2018: a review of 3449 articles in leading orthopaedic journals. *Patient Saf Surg* **14,** 23 (2020). https://doi.org/10.1186/s13037-020-00246-6

2009 - Level of Evidence Evaluation

- Thirty-eight residents from 5 orthopedic surgery training programs, from year-in-training 3 to 5, determined the levels of evidence rating of 10 blinded articles representing all levels of evidence types in the orthopedic literature.

- Residents graded the level of evidence correctly in fewer than half the papers. These findings indicate that resident knowledge of levels of evidence criteria is limited and suggest a need for more education in this area.

Analysis of Orthopedic Resident's Ability to Apply Levels of Evidence Criteria to Scientific Articles

- 25 U.S. orthopedic residents and 15 4th year medical students interviewing for a residency position in orthopedic surgery were provided with the article title, the abstract, and the complete methods section for 15 articles from the American Volume of Journal of Bone and Joint Surgery
- The assigned LoE designation was withheld but access to the LoE criteria used by Journal of Bone and Joint Surgery was provided.
- Each participant was asked to assign a study type and LoE designation for each article.

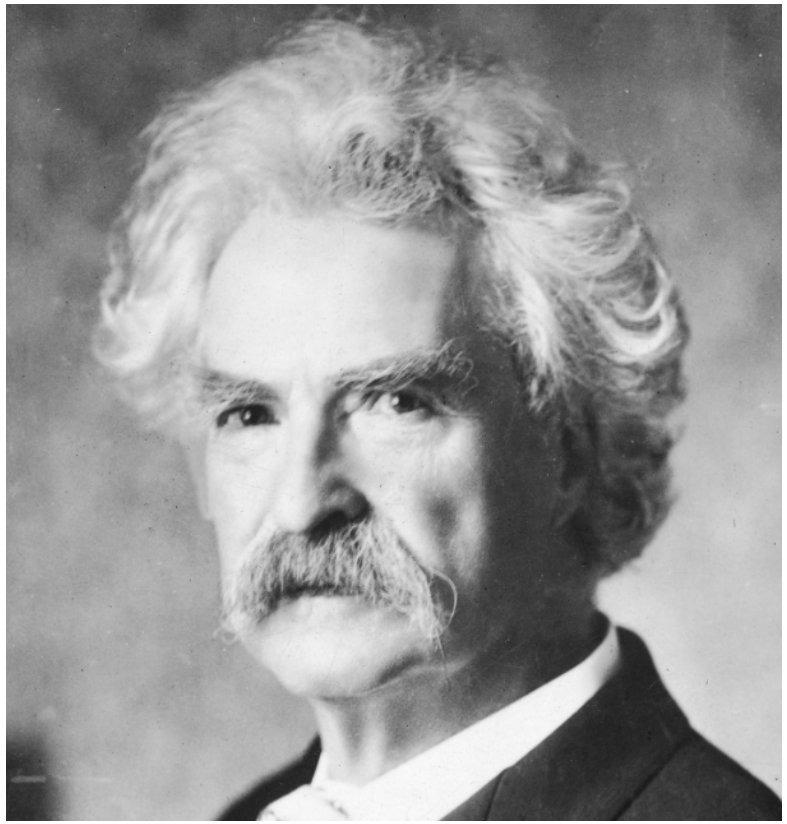
Grandizio et al. J. Surg Education May-Jun 2016;73(3):381-5.doi: 10.1016/j.jsurg.2015.11.012. Epub 2016 Jan 28.

Analysis of Orthopedic Resident's Ability to Apply Levels of Evidence Criteria to Scientific Articles

- There were more correct responses regarding the study type (67%) than for LoE designation (39%).
- The percentage of correct responses for study type and LoE increased with more years of training ($p = 0.005$ and $p = 0.002$).
- Residents had a higher proportion of correct LoE responses overall than medical students, but this difference did not reach statistical significance (42% vs. 35%, $p = 0.07$).
- Strategies to improve resident understanding of LoE guidelines need to be incorporated into orthopedic residencies, especially when considering the increased emphasis on evidence-based medicine.

Grandizio et al. J. Surg Education May-Jun 2016;73(3):381-5.doi: 10.1016/j.jsurg.2015.11.012. Epub 2016 Jan 28.

Mark Twain

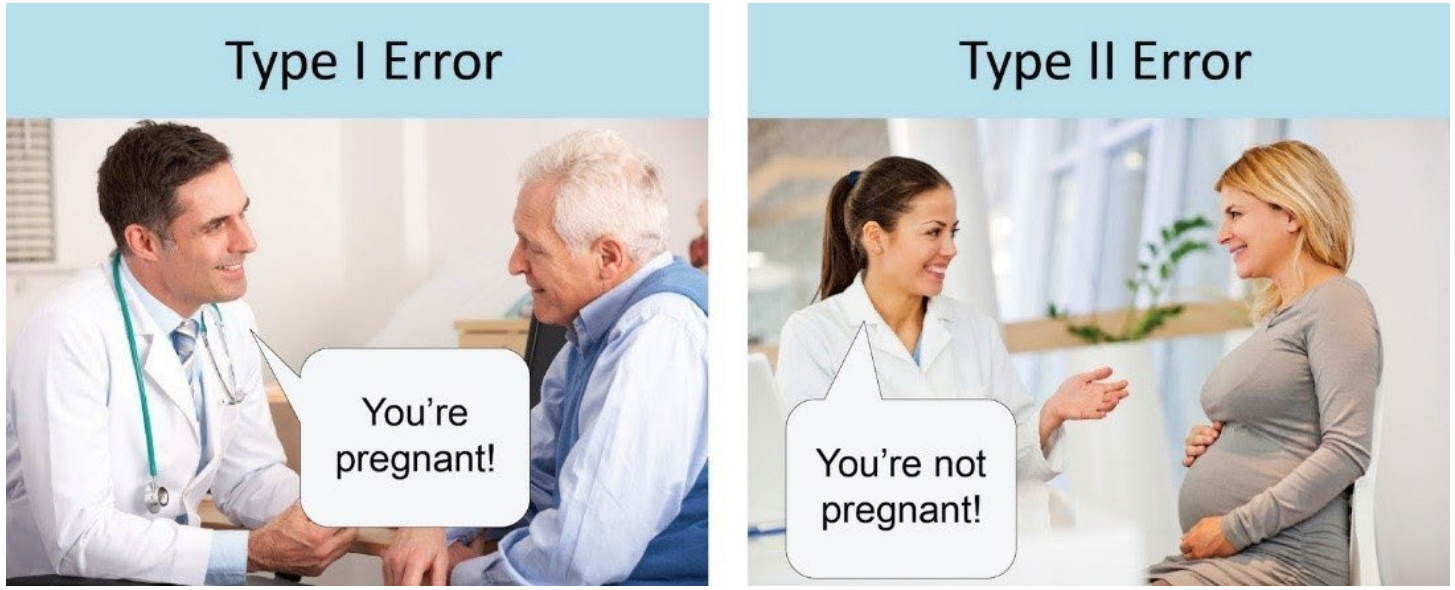


"There are lies, there are damn lies, and there are statistics"

Paraphrased from Robert Giffen, President Statistical Society 1882-1884

P Value: what does it mean?

- $p \leq .05$ means we are 95% sure the observed difference is not by chance



False positive

False negative

Bhandari et al. The risk of false-positive results in orthopaedic surgical trials. CORR 2003 Aug;(413):63-9

Kocher MS, Zurakowski D. Clinical Epidemiology and Biostatistics: A Primer for Orthopaedic Surgeons. JBJS 2004:86:607-620

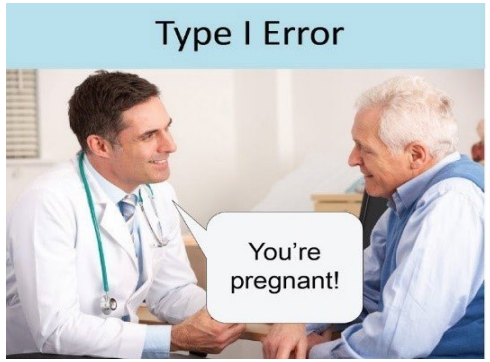
Study Power

- The power of a study is the probability that it will demonstrate a difference between two treatments when one actually exists. Power (1 - β) is simply the complement of the type-II error.
- If we accept a 20% chance of an incorrect study conclusion (beta; = 0.20), we also accept the corollary that we will come to the correct conclusion 80% of the time.
- Study power can be postulated before the start of a clinical trial to estimate the optimal sample size or it can be evaluated after the completion of a study to determine whether the negative findings were true or more likely due to chance.

Lochner, Heather V. MSc; Bhandari, Mohit MD. MSc; Tornetta, Paul III MD. Type-II Error Rates (Beta Errors) of Randomized Trials in Orthopaedic Trauma, The Journal of Bone & Joint Surgery: November 2001 - Volume 83 Issue 11 - p 1650-1655

Type I "*Alpha*" Error

- Incorrectly rejecting the "null hypothesis" when it is actually true (False Positive)
- The "Power" of the test
- Conventionally adequate Power is ≥80% to detect a difference associated with a "significant" treatment effect
- Study needs to have an adequate number (n) to detect a difference in treatment effect greater than 80% of the time



False positive

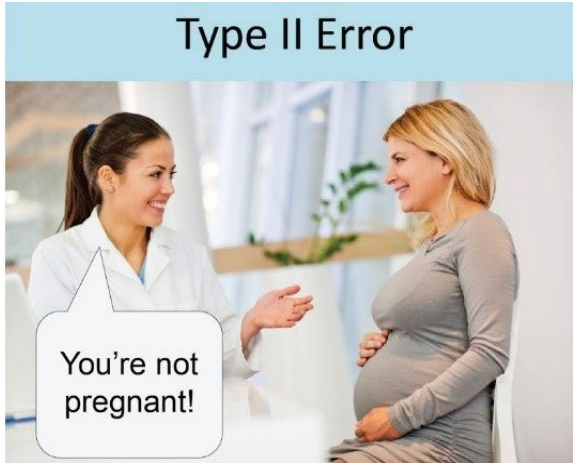
Type 1 Error - Multiple Testing

- 127 articles in two major orthopaedic journals analyzed statistically
- Multiple instances of uncorrected multiple outcome testing causing an "estimated median risk of obtaining at least one significant result for uncorrected studies was calculated to be 54% for both journals "

Wallenkamp et el. Multiple Testing in Orthopaedic Literature: a common problem. BMC Res Notes. 2013; 6: 374.

Type II "*Beta*" Error

"The results of randomized studies are given much greater weight than are those of retrospective or case-controlled studies, but if they are underpowered they can lead to conclusions that may justify an inferior treatment."



False negative

Lochner, Heather V. MSc; Bhandari, Mohit MD. MSc; Tornetta, Paul III MD. Type-II Error Rates (Beta Errors) of Randomized Trials in Orthopaedic Trauma, The Journal of Bone & Joint Surgery: November 2001 - Volume 83 Issue 11 - p 1650-1655

Type II "*Beta*" Error

- Analyzed 117 studies in which a total of 19,942 patients with orthopaedic trauma had been randomized. Sample sizes ranged from ten to 662 patients
- Majority (34%) of trials involved the treatment of hip fractures.
- The mean study power among the 117 trials was 24.65% (range, 2% to 99%).
- The type-II error rate for primary outcomes was 90.52%.

Lochner, Heather V. MSc; Bhandari, Mohit MD. MSc; Tornetta, Paul III MD. Type-II Error Rates (Beta Errors) of Randomized Trials in Orthopaedic Trauma, The Journal of Bone & Joint Surgery: November 2001 - Volume 83 Issue 11 - p 1650-1655

Type II "*Beta*" Error

"Conclusions: Mean type-II error rates in the orthopaedic trauma trials that we analyzed exceeded accepted standards. Investigators can reduce type-II error rates by performing power and sample-size calculations prior to conducting a trial."

Lochner, Heather V. MSc; Bhandari, Mohit MDMSc; Tornetta, Paul III MD Type-II Error Rates (Beta Errors) of Randomized Trials in Orthopaedic Trauma, The Journal of Bone & Joint Surgery: November 2001 - Volume 83 - Issue 11 - p 1650-1655

P Value

- Not the be all and end all
- “Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.”
- “Proper inference requires full reporting and transparency.”
- **“A p-value, or statistical significance, does not measure the size of an effect or the importance of a result”.**

American Statistical Society 2016

Confidence Interval (CI)

- "Clinical" Importance vs "Statistical" Significance
- Clinical importance reflects how big an effect is noted
- Statistical Significance can only imply whether there is a difference between groups being compared, not how much
- Confidence intervals *combine* the analyses of effect size and statistical significance
- Actual degree of difference between treatment groups is often more important than whether it was statistically significant

Confidence Intervals

- CI is the range of values in which we are “fairly sure” our true value lies
- Calculated from the mean and standard deviation

$$X \pm Z\frac{s}{\sqrt{n}}$$

X = Mean observed value

Z = CI table value for desired CI

S = standard deviation

N = number of observations

Confidence Interval	z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

Confidence Interval

- 90% Confidence Interval means that there is 90% likelihood of a studied treatment's true therapeutic effect will lie within the range of outcome values seen.
- When two 95% CI's for different treatments do not overlap, or when a certain observed value for one treatment is not contained within the other groups range of values, we are sure that there is a statistical difference at the 5% level
- More clinically informative than a study narrowly disproving the "null hypothesis.
- Unfortunately, CI's are not typically reported in clinical research

Confidence Interval: Reporting

- 2009 Vavken etal. 8 Ortho Journals RCT's 2000,2003,2006

  22% of 88 randomly chosen RCT's reported CI's

- 2020 Raittio & Reito 8 Ortho Journals reviewed 2016-2017

  19% of 160 RCT's had CI's

"It was also a concern to find that only one-fifth of studies reported the confidence intervals for mean difference value, which is in line with the situation in orthopedics a decade ago"

Vavken P, et. al. The use of confidence intervals in reporting orthopaedic research findings. Clin Orthop Relat Res 2009; 467(12): 3334–9

Lauri Raittio & Aleksi Reito (2020) Assessing variability and uncertainty in orthopedic randomized controlled trials, Acta Orthopaedica, 91:4, 479-484, DOI: 10.1080/17453674.2020.1755932

P “Hacking”: Selective Reporting

In order to establish a p Value ≤ 0.05 investigators may “clean up data” with techniques to reach publication level “Significance”

1. Collecting many response variables and deciding what to report on given statistical “post hoc” analysis
2. Deciding whether to include or exclude study subject outliers post analysis
3. Split treatment groups after treatment result analysis

Scientist behaving badly

US scientists engage in a range of behaviors extending far beyond falsification, fabrication and plagiarism

Examples:

- overlooking others use of flawed data or questionable interpretation of data

- changing the design, methodology or results of a study inn response to pressure from a funding source or poor study design

- circumventing certain minor aspects of human subject requirements

Martinson et. al. Nature 435 (9)2005

Scientist behaving badly

US scientists engage in a range of behaviors extending far beyond falsification, fabrication and plagiarism

Examples:

- inappropriate assignment of authorship credit

- inappropriate or inadequate research design

- inadequate record keeping

- dropping observations or data points from analyses based on gut feeling that they were inaccurate

Martinson et. al. Nature 435 (9)2005

“Beware of “Misbehavior”

- Hypothesis: ≥85% of literature examined would report supported hypotheses
- 215 Original Research Articles 1/1/2019-5/31/2019
- Published in 4 Journals British Journal of Sports Medicine, Sports Medicine, American Journal of Sports Medicine ,Journal of Orthopaedic and Sports Physical Therapy
- Only 129 (60%) reported at least one study hypothesis
- 106 (82.2%) reported the primary hypothesis was supported by the results
- “Notably, the proportion of Registered Reports reporting statistically significant results in support of study hypotheses, following data collection and analysis, is approximately 40%”[1]
- Impossible to accurately assess % supported hypotheses as 40% did not have a pre-established experimental hypothesis

Beware of “Misbehavior”

- Questionable research practices (QRPs) are intentional and “unintentional practices that can occur when designing, conducting, analyzing, and reporting research, producing biased study results.”

- “Academic culture prioritizes novel research and positive, rather than negative, study findings.”

- “One third of scientists admit to using QRP’s such as P-Hacking, selective outcome reporting and hypothesizing after the results are known (*Hark*ing) to generate statistically significant results.”[1]

Fanelli,D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One2009;4:e5738.doi;10.1371/journal.pone.00057pmid:http://www.ncbi.nlm.nih.gov/pubmed/19778950

Discerning Truth from Fiction

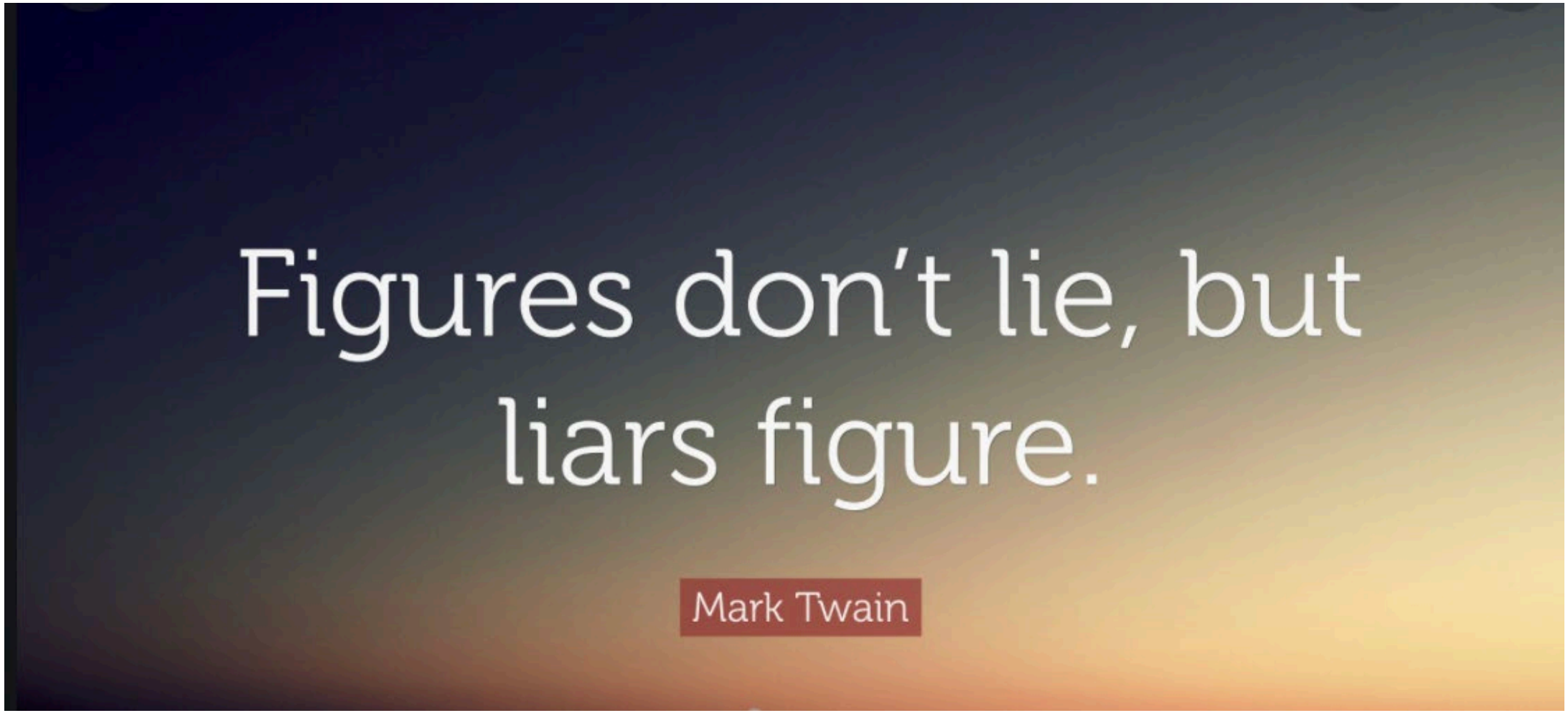
"The determinants of the truth of a knowledge claim lie in combination of evidence both within and outside a given experiment, including the plausibility and evidential support of the proposed underlying mechanism. If that mechanism is unlikely, as with homeopathy or perhaps intercessory prayer, a low *P* value is not going to make a treatment based on that mechanism plausible"

S. S. Leopold, Editor-in-Chief, *Clinical Orthopaedics and Related Research®*

Threshold P Values in Orthopaedic Research—We Know the Problem. What is the Solution?

Goodman S. A dirty dozen: Twelve p-value misconceptions. Seminars in Hematology. 2008;45:135–140.

Mark Twain: Remember



Quick critical appraisal checklist

Item	
1. Was there a clear, focused and answerable study question?	
2. Were patients allocated randomly to trial groups?	
3. Were patients similar at baseline in terms of demographics and comorbidity?	
4. Did the authors perform a sample size calculation?	
5. Was there any blinding?	
6. Were patients treated equally apart from the study interventions?	
7. Were patients analyzed as randomized (intent to treat)?	
8. Do the authors provide sufficient numerical information to recalculate the results?	

You are the ultimate “reviewer”

- Determine What you need to read
- Read Regularly
- Skim articles in a select journal group
- Choose those of interest
- Look for BIAS
- Assess External Validity
- STUDY CAREFULLY METHODS and RESULTS
- Assess Internal Validity
- Does Data support the author’s conclusions?