# How to Critically Review Journal Articles and Understand Their Statistics

Chad A Krueger, MD

Created July 2016

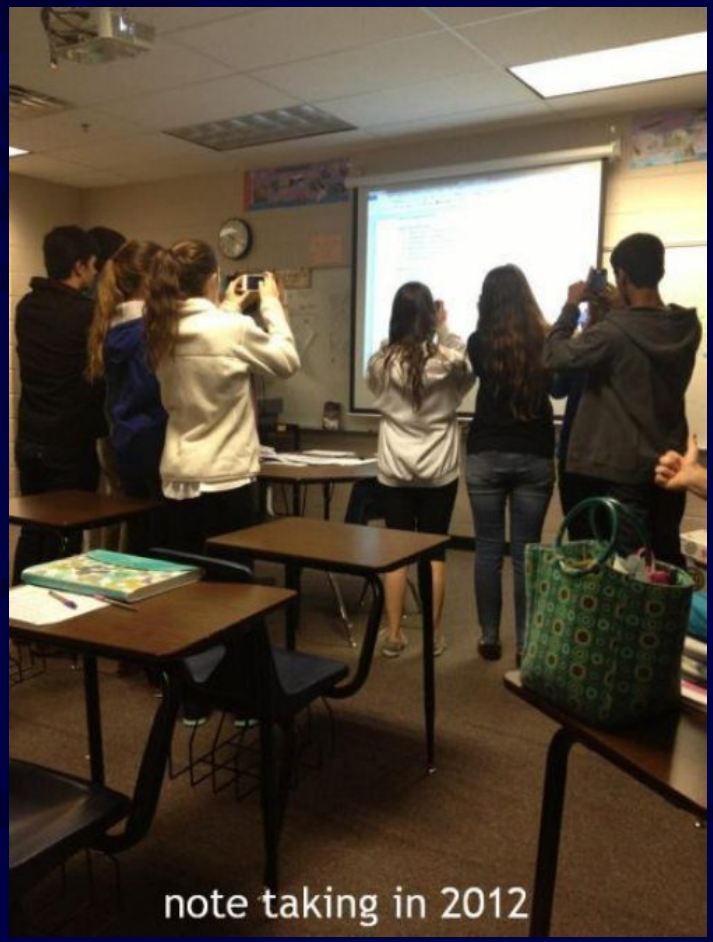
# Why do we need to be taught how to read?

- Many orthopaedic surgeons receive little formal instruction on how to evaluate educational material
- Over 12,000 articles published in orthopaedic surgery or sports medicine in 2013 alone
- There is not enough time to read all of the literature
  - Must determine what is important to read and then learn how to read it

Krueger et al. What to Read and How to Read It: A Guide for Orthopaedic Surgeons. JBJS. 2016. 98:243-9

# Why do we need to be taught how to read?

- ‘Keeping up’ with literature no longer possible
  - 2-4,000 citations added to MEDLINE daily
- Can you honestly say that you understand what most journal articles are saying?




Gillespie et al. Clin Orthop Relat Res 2003;413:133-45
Clough et al. Inst Course Lecture. 2011;60:607-618

# What is the primary type of evidence orthopaedic surgeons use in clinical decisions?

- Almost 50% of American Orthopaedic Association members rely on personal experience or expert opinion for decision making
  - Read: Level 5 evidence
- Less than 15% rely on randomized control trials
  - Read: Level 1 or 2 evidence

# Why is EBM not universally embraced by orthopaedic surgeons?

- Over 75% of orthopaedic surgeons feel that EBM does not relate to their practice and/or they don't believe the published data

Schemitsch et al. JBJS Am. 2009;91:1264-73

# That was orthopaedic attendings, what about residents?

- Only 28% of surgery residents feel they have enough training to properly incorporate EBM
- Many residents who don't feel comfortable evaluating literature become attendings who do not properly use EBM in their decision making
- There is likely a historical basis to this problem

Krueger et al. What to Read and How to Read It: A Guide for Orthopaedic Surgeons. JBJS. 2016. 98:243-9

# Evidence based medicine

- Evolved from epidemiology
- Canadian medical association journal – 1980
  - ‘New teaching of technique’
- “Evidence-based Medicine”- American College of Physicians Journal club 1991
- The idea of EBM is only 25 years old

Dr. David Sackett

Dr. Gordon Guyatt

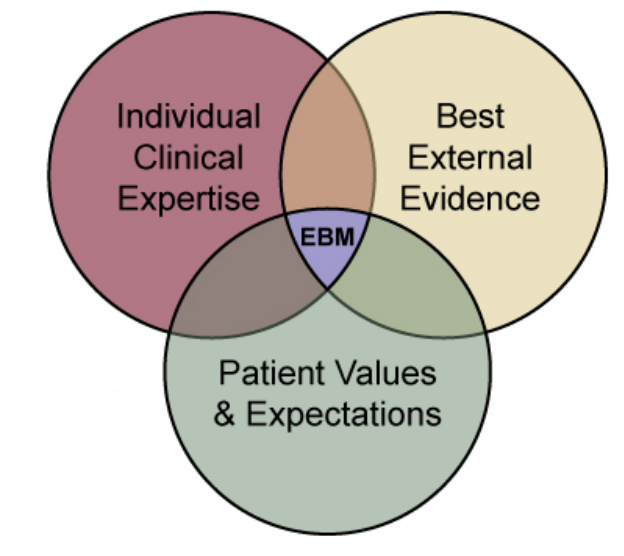
Hoppe et al. JBJS Am. 2009;91:2-9
Hurwitz et al. JBJS Am. 2006: 88;1873-9
http://www.uic.edu/depts/lib/lhs/resources/guides/ebmonline/EBM_Intro_revised2/EBM_Intro_revised2.html

# Evidence based medicine

- Shift away from apprenticeships
- Different from ‘lineage based knowledge’
- Focuses on best practices, not where trained
- Collective decision of what is best
  - Decrease cognitive and research biases

Hoppe et al. JBJS Am. 2009;91:2-9
Hurwitz et al. JBJS Am. 2006: 88;1873-9

http://www.uic.edu/depts/lib/lhs/resources/guides/ebmonline/EBM_Intro_revised2/EBM_Intro_revised2.html



# Levels of Evidence

- Only 11.3% of all orthopaedic articles published are of the level 1 variety

Obremsky WT et al.Level of evidence in orthopaedic journals. JBJS Am. 2005;87:2632-8

- Some questions are impossible to study using level 1 evidence due to ethical and other constraints
  - RCTs are also not needed if the effect of an intervention is dramatic or when the possibility of confounding variables can be ignored
    - Do we need a level 1 study to show that anesthesia during surgery improves patient outcomes?
- These levels are not absolute
  - there can be great level 4 studies and poor level 1 studies

# Levels of evidence

- Level 5
  - Expert opinion
    - Case report
    - Personal observation
    - “I recommend treatment *x* because when I do treatment *x* it works well.”
- Level 4
  - Case series
    - There is no control group
      - Prognostic or Diagnostic studies
    - The reference standard is poor
      - Diagnostic studies
    - Minimal sensitivity analysis
      - Economic studies

Karlsson et al. A Practical Guide to Research: Design, Execution and Publication. Arthroscopy,2011;27:S1-S112

# Levels of evidence

- Level 3
- Therapeutic and Diagnostic studies
  - Case-control
    - Compare patients with a disease or treatment to those without
  - Retrospective cohort
    - Compare patients who received treatment or disease exposure prior to the start of the study
  - Nonconsecutive patients
  - Inconsistently applied ‘gold standard’
- Systematic review or meta-analysis of Level 3 studies

Karlsson et al. A Practical Guide to Research: Design, Execution and Publication. Arthroscopy,2011;27:S1-S112

# Levels of evidence

- Level 2
- Therapeutic and Diagnostic studies
  - Prospective cohort
  - Lesser RCTs (<80% follow up, no blinding)
  - Consecutive patients against gold standard for every patient
- Prognostic studies
  - Untreated controls from RCT
  - Retrospective
- Systematic review or meta-analysis of Level 2 studies

Karlsson et al. A Practical Guide to Research: Design, Execution and Publication. Arthroscopy,2011;27:S1-S112

# Levels of evidence

- Level 1
- Therapeutic, Prognostic and Diagnostic studies
  - Randomized Controlled Trials
  - Blinding not necessary
  - Proper statistical analysis
- Economic studies
  - Multiway sensitivity analysis
- Systematic review or meta-analysis of Level 1 studies

Karlsson et al. A Practical Guide to Research: Design, Execution and Publication. Arthroscopy,2011;27:S1-S112

# Types of reading

- Knowledge reading
  - Learning of a subject
  - Review articles
- Apply-to-practice reading
  - Specific questions
  - Original research articles
- Immediate-knowledge reading
  - Case-based reading

# Resident reading

- Textbooks
- Review articles- (eg Journal of the American Academy of Orthopaedic Surgeons)
- Orthopaedic Knowledge Update and Orthopaedic Knowledge Update Trauma books
  - Annotated bibliography
    - Provide good overview of important articles
  - Information vetted by subject matter experts
- Specific scientific articles
  - Occasionally
  - Need understanding before interpretation

# Attending reading

- Generalist
  - JBJS Am
  - Areas of interest/specific questions
  - Areas of weakness (similar to resident reading)
- Traumatologist
  - Journal of Orthopedic Trauma
  - Relevant Journal of Bone and Joint Surgeons American volume articles
  - Areas of weakness (similar to resident reading)
    - JAAOS, OKUs, textbooks

# The Reading Pyramid

- Junior residents
  - Focus on gathering objective data
- Senior residents
  - Comparing data to gather information
- Junior attendings
  - Developing knowledge by adding experiences to their information
- Senior attendings
  - Reflecting on their experiences, growing wisdom

Krueger et al. What to Read and How to Read It: A Guide for Orthopaedic Surgeons. JBJS. 2016. 98:243-9

# How to improve your reading

- Have a clear goal in mind while reading
  - What are you trying to learn?
  - Increases focus and retention
  - Provides framework for determining external validity and potential conclusions
- The reader must be aware of his or her biases prior to reading
  - These influence the interpretation of the data

Krueger et al. What to Read and How to Read It: A Guide for Orthopaedic Surgeons. JBJS. 2016. 98:243-9

# How to improve your reading

- If reading for specific knowledge
  - Scan an article or text until that information is found
- If reading for general knowledge
  - Read information from start to finish
  - Provides more context to help increase associations and retention
- The less familiar a reader is with a topic, the more basic the text should be
  - If the reading structure is too complex, the reader will have a hard time understanding the information it contains

Krueger et al. What to Read and How to Read It: A Guide for Orthopaedic Surgeons. JBJS. 2016. 98:243-9

# Book chapters

- Very detailed
- Time consuming
- Skimming may provide a better understanding of the general content because it limits details to a manageable level
- When skimming
  - Most important data can be found in tables or figures
  - Actively determine how content fits with your current knowledge

Krueger et al. What to Read and How to Read It: A Guide for Orthopaedic Surgeons. JBJS. 2016. 98:243-9

# Reading a Book Chapter

- Introduction
  - What is the topic? What are you trying to learn?
  - Does the author have any potential financial or intellectual biases?
  - Read first two paragraphs
- Body
  - Read first two and last sentence of each paragraph
  - Read all tables, figures and diagrams and determine main conclusions
- Conclusion
  - Read conclusion
  - Think of how the chapter fits with your current knowledge

Krueger et al. What to Read and How to Read It: A Guide for Orthopaedic Surgeons. JBJS. 2016. 98:243-9

# Journal Articles

- Skimming is not recommended
  - Key details determine internal and external validity of an article
- External validity easiest to determine first
  - Does the article apply to your practice?
  - If not, move on
- Internal validity
  - Are the methods of the study reproducible and likely to provide unbiased results
  - If not, the results are invalid

Krueger et al. What to Read and How to Read It: A Guide for Orthopaedic Surgeons. JBJS. 2016. 98:243-9

# Reading a Journal Article

- Title and Abstract
  - Determine external validity
  - Does the paper apply to your practice?
- Evaluate Methods
  - What are the methodological flaws? (internal validity)
  - Do they invalidate the study?
- Results/Figures and Tables
  - Are the results interesting/compelling?
  - Are the results clinically relevant?
- Read the entire paper
  - Do the conclusions mesh with the results?
  - Are there conflicts of interest that may bias the results or conclusions?
  - How does the study fit with your current knowledge?

Krueger et al. What to Read and How to Read It: A Guide for Orthopaedic Surgeons. JBJS. 2016. 98:243-9

# So you found an article…

- What type of study is it?
  - Observational
    - No intervention given, observing outcomes
  - Experimental
    - Provide intervention, measure outcome
    - Retrospective, prospective
- What is the article about?
  - Observational study
  - Article about a prognosis
  - Therapeutic treatment
  - Diagnostic test
  - Meta-analysis
- What are you trying to learn from it?

# Determining what you are reading

- Case-control
  - Compares with an outcome to those without the same outcome
    - Allows an odds ratio to be calculated
- Cross-sectional surveys
  - Determine prevalence of condition at a specific time
- Prospective cohort studies
  - Can be used to determine incidence

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

# Observational Studies

- Two common goals of observational studies
  - Describing the likelihood of a certain outcome
  - Providing an association between a diagnosis or a treatment and a condition
- Extremely useful in helping to develop hypotheses for future research
  - For example- it was observational studies looking a clavicular nonunions that lead to the prospective studies comparing nonoperative and operative fixation
- 3 main types of observational studies
  - Cohort
  - Case series
  - Case reports

Hoppe DJ *et al.* Hierarchy of Evidence: Where Observational Studies Fit in and Why We Need Them. JBJS 2009;91 s3.2-9

# Observational Studies

- Observed differences may be the result of a confounding variable
  - A variable that relates to both the dependent and independent variable
- Without control subjects, it is very hard to account for these confounding factors
  - Gamblers are more likely to smoke, smoking causes cancer, so gamblers are more likely to get cancer
  - But gambling does not cause cancer directly

Hoppe DJ *et al.* Hierarchy of Evidence: Where Observational Studies Fit in and Why We Need Them. JBJS 2009;91 s3.2-9

# Bias and Confounding

- Bias
  - An inclination or prejudice for or against one group
  - Leads to results that are not ‘true’
- Confounding
  - Confusing associations and effects from extraneous variables with those variables studied

# Types of bias

- Attrition bias
  - Dissimilar groups of patients lost to follow up
- Expertise bias
  - One group of patients has a surgeon who has more expertise than another
- Recall bias
  - Subjects remembering exposures/treatments in a nonuniform manner
- Selection bias
  - Dissimilar patients comprise the different groups being compared
- Information
  - Bias resulting from measurement error or data misclassification

# Types of bias

- Verification bias- only test reference standard for those with positive test, assume those with negative test don’t have target condition
  - When the test is invasive, surgeons less likely to test it when disease probability is low
- The test is also dependent on the people conducting the test
  - There may be variability within this group that could lead to poor external validity

# Bias and Confounding

- Methods to decrease confounding
  - Matching
  - Stratification
  - Multivariable regression
- Factors contributing to bias
  - Missing data
    - If data is missing randomly, only decreases power
    - If nonrandomly missing, biases the findings
      - No way to offset this biases through computation

Morshed S, *et al*. Analysis of Observational Studies: A Guide to Understanding Statistical Methods. JBJS 91;50-60 2009

# Observational studies

- Observational ≠ not usable
  - Letournel and Judet, McKee clavicle studies
- Not ideal for therapy
  - You may not have the same skill as Letournel for acetabluar fractures and cannot expect the same outcomes
- Best for prognostic factors, natural history, adverse events or unethical studies
  - Smoking on fracture healing
  - Contamination in open wounds leading to infection

Hoppe et al. JBJS Am.Hierarchy of Evidence: Where Observational Studies Fit in and Why We Need Them. 2009;91:2-9

# Observational studies

- Observational studies may be associated with larger positive treatment effects than randomized trials
  - It may show that a certain treatment or therapy has a greater effect on an outcome than it does in reality
- However, some studies have shown no differences in results obtained by observational studies and the results found from RCT

Hoppe DJ *et al.* Hierarchy of Evidence: Where Observational Studies Fit in and Why We Need Them. JBJS 2009;91 s3.2-9

# What makes a case series good?

- Subjects that represent the study population well
- Reproducible intervention
- Clinically important outcome measures
- As much follow up as possible
- Basic statistical analysis
  - Rate, risk, confidence intervals

Morshed S, *et al*. Analysis of Observational Studies: A Guide to Understanding Statistical Methods. JBJS 91;50-60 2009

# Articles relating to therapy

- If the prognostic factors are not balanced between treatment groups, the outcomes will be biased
  - This is why observational trials tend to show larger treatment effects than RCTs- RCTs have randomized treatment groups
  - Importance of ensuring groups are randomized and similar
  - Check to see if the prognostic factors for each group are listed and similar

Bhandari et al. Users' guide to the orthopaedic literature: how to use an article about a surgical therapy. JBJS Am. 2001;83:916-26

# What makes a good article relating to prognosis?

- Is the population similar for both groups or similar to your own practice?
  - Mortality rate at tertiary center vs community hospital
- Are control/treatment groups similar?
- Randomization or matching of the study groups?

Bhandari et al. User's Guide to the Orthopaedic Literature: How to Use an Article about Prognosis. JBJS Am. 2001;83:1555-64

# What makes a good articles relating to prognosis?

- Are the diseases of similar severity?
  - Stage III/IV cancer pts versus cancer pts who died
  - Operative delay with hip fractures
    - More than 3 days ➔ increased mortality
    - Adjust for pre-existing conditions using ASA ➔ no difference
    - Illness severity, not delay in treatment most important
- What is the external validity?
  - Does the study relate to your practice specifically?

Bhandari et al. User's Guide to the Orthopaedic Literature: How to Use an Article about Prognosis. JBJS Am. 2001;83:1555-64

# What makes a good articles relating to prognosis?

- Follow up
  - Was the follow up of sufficient length?
- Are those lost to follow up likely to have different outcome than those not lost?
  - Trauma patients lost doing just as well as those in clinic?
- Outcome criteria
  - Was it standard to all subjects?
  - Were evaluators blinded?

Bhandari et al. User's Guide to the Orthopaedic Literature: How to Use an Article about Prognosis. JBJS Am. 2001;83:1555-64

# About lost to follow up…

- Outcome
- Lost to follow-up → compromised validity
  - Rules of thumb (20% or less) inaccurate
- Assume a worst-case scenario for lost to follow-up
  - If it does NOT change treatment effect, okay
  - If it does change change treatment effect, problem

Bhandari et al. User's Guide to the Orthopaedic Literature: How to Use an Article About a Surgical Therapy. JBJS Am. 2001;83:916-26

# What to look for in a study about a diagnostic test

- Is there diagnostic uncertainity?
  - When severely diseased subjects are compared to healthy subjects, there is an overestimation of test performance
  - This makes it much less clear if the test is useful for patients who are in the ‘gray zone’ where the test is most likely to be needed
- The test should be tested on patients who are most likely to need the test
- The test result should be compared to an independent, gold standard test so that the results of the new test are not biased

Bhandari et al. User’s Guide to the Orthopaedic Literature: How to Use an Article about a Diagnostic Test. JBJS Am. 2003;85:1133-40

# What to look for in a study about a diagnostic test

- Study groups need to be of similar disease to isolate the test performance compared to the gold standard
  - Healthy volunteers vs diseased individuals overestimate test performance threefold
- Need patients with low, high and moderate suspicion of disease
  - Allows determination if test is valid for all groups

Bhandari et al. User's Guide to the Orthopaedic Literature: How to Use an Article about a Diagnostic Test. JBJS Am. 2003;85:1133-40

# A word about meta-analyses

- Large increase in number published
  - 5 fold increase from 1999 to 2008
  - Over half of all meta-analyses published in 2005 and 2008 had methodological flaws
    - 30% in 2008 had major methodological flaws
- 50-60% of meta-analyses have methodological flaws
- Difference between meta-analysis and systematic review
  - Review- summary of medical literature addressing a focused clinical topic
  - Meta-analysis- a systematic review that uses statistical analysis to summarize the results
- The results of a meta-analysis are only as good as the evidence include within their evaluation

Dijkman et al Twenty years of meta-analyses in orthopaedic surgery: has quality kept up with quantity? JBJS Am. 2010;92:48-57

# Checklist for determining quality of RCT

- Was the generation of allocation sequences adequate?
- Was the treatment allocation concealed?
- Were details of the intervention of each group explained?
- Did providers in each group have enough skill?
- Was patient adherence monitored?
- Were participants blinded?
- Were providers blinded?
- Were outcome assessors blinded?
- Was the follow up the same for each group?
- Were the outcomes analyzed according to the ‘intention to treat’ principle?

Jones J, Hunter D. Consensus methods for medical and health services research. BMJ. 1995. 311:376-80

# How to determine what articles to read?

- It starts with external validity
- Is your patient population the same as the study's?
- Do you perform the same type of treatment?
- Is your experience similar to that of the author?
- How is your practice different than that studied?
- In short- can the results of the study apply to your practice?

# If the study relates to your practice

- The next step is determining if the study is methodologically sound
- Internal validity
- Are the methods sound or do they invalidate the study results?
- Go to the methods section

# Internal validity

- Do the methods make sense?
- Is there bias imbedded in the study?
- Are their ‘catastrophic failures’ that make study invalid
  - Often not in abstract
- Just because its in a ‘good’ journal…



## The Role of the Acetabular Labrum and the Transverse Acetabular Ligament in Load Transmission in the Hip*

BY GREGORY A. KONRATH, M.D.†, ANDREW J. HAMEL, B.B.S.‡, STEVE A. OLSON, M.D.§, BRIAN BAY, PH.D.§, AND NEIL A. SHARKEY, PH.D.‡, SACRAMENTO, CALIFORNIA

# Internal validity

- Conclusion- labrum doesn't affect hip stability
- In the methods the study discusses how the joint capsule removed and…

The moistened femoral head was covered with a thin layer of latex (a Trojan latex condom; Carter Wallace, New York, N.Y.). Two layers of prescale film were cut into a star shape and were applied to the latex with photographic mounting adhesive. A second layer of latex then

# Internal Validity

- So this study is running a test on the biomechanics of the hip with the assumption that a latex condom and fuji film had the same biomechanical properties as the labrum
- This likely invalidates the results

# Internal validity

- Better?

An in vitro investigation of the acetabular labral seal in hip joint mechanics

S.J. Ferguson[a,b,*], J.T. Bryant[c], R. Ganz[d], K. Ito[b]

[a] *M.E. Müller Institute for Biomechanics, University of Bern, Murtenstrasse 35, Postbox 30, CH-3010 Bern, Switzerland*
[b] *AO Research Institute, Davos, Switzerland*
[c] *Department of Mechanical Engineering, Queen's University, Kingston, Canada*
[d] *Department of Orthopaedic Surgery, University of Bern, Inselspital, Bern, Switzerland*

Accepted 04 November 2002

- Hip capsule with joint fluid
  - Integrity of natural joint intact
- No condoms or fugi film
- Conclusion- labrum plays a role in cartilage compaction at the hip

# Internal validity

- Are the methods consistent
- Questions: ORIF vs arthroplasty for proximal femur fractures
- Level 1 study in JBJS stated- arthroplasty best
- When reading the methods- residents did all of the ORIFs, attendings did the arthroplasty
- Did arthroplasty do better because it was the attending doing the surgery?
  - The interpretation of the results should be cautious

Blomfeldt R, et al. Comparison of Internal Fixation with Total Hip Replacement for Displaced Femoral Neck Fractures. J Bone Joint Surg. 2005. 87; 1680-88

# If the study seems valid

- Results next
  - Need to interpret for yourself
  - Step back- do they make sense?
- Look at figures and tables first
  - Most important data contained here
  - Be clear of what you are looking at
    - Anything can be graphed
- Tables
  - Provide clear data
  - Harder to interpret trends and recognize outliers

# Results

- Must form your own conclusions about the results
- Use this interpretation to read the discussion/conclusion
- Results-conclusion mismatch
  - Results show *x*, conclusion states *y*


MILITARY MEDICINE, 172, 7:777, 2007

Application of the Mangled Extremity Severity Score in a Combat Setting

# Result-conclusion mismatch

- Results showed that 6 and 7 patients, respectively, in both groups had MESS of 7 +/- 2
- The means of each group were different but those were filled by patients on each extreme
- Those extreme patients are not where the controversy lies in terms of salvage or amputation

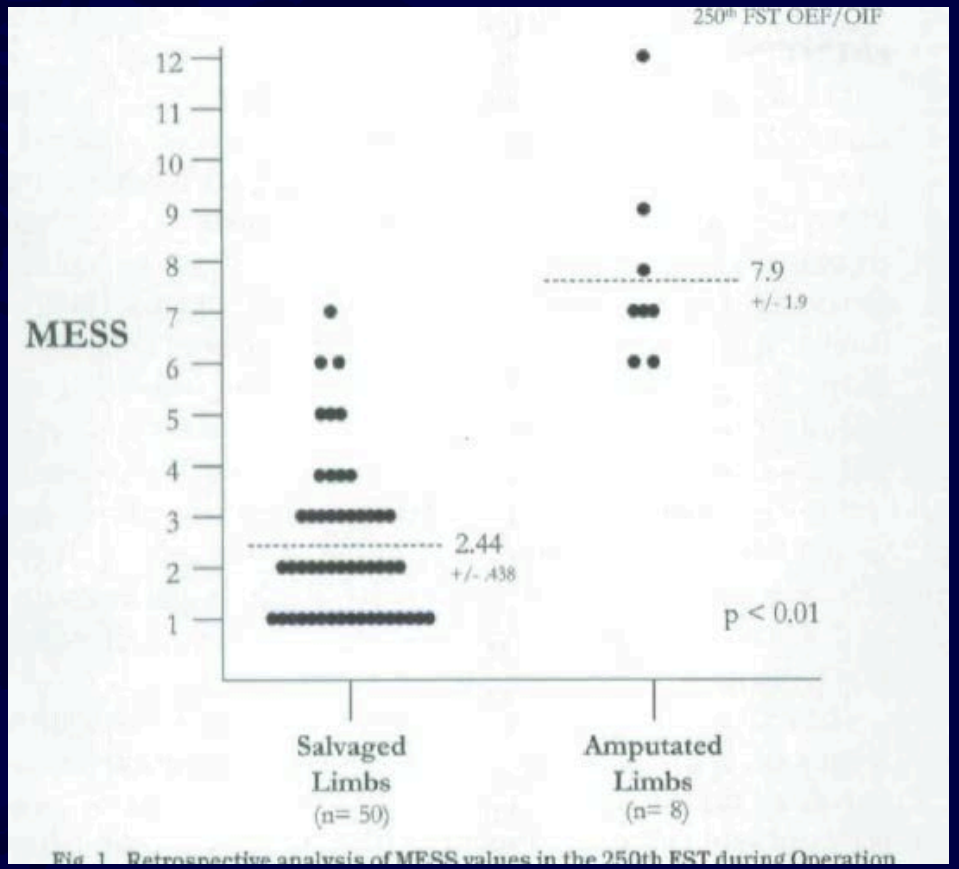
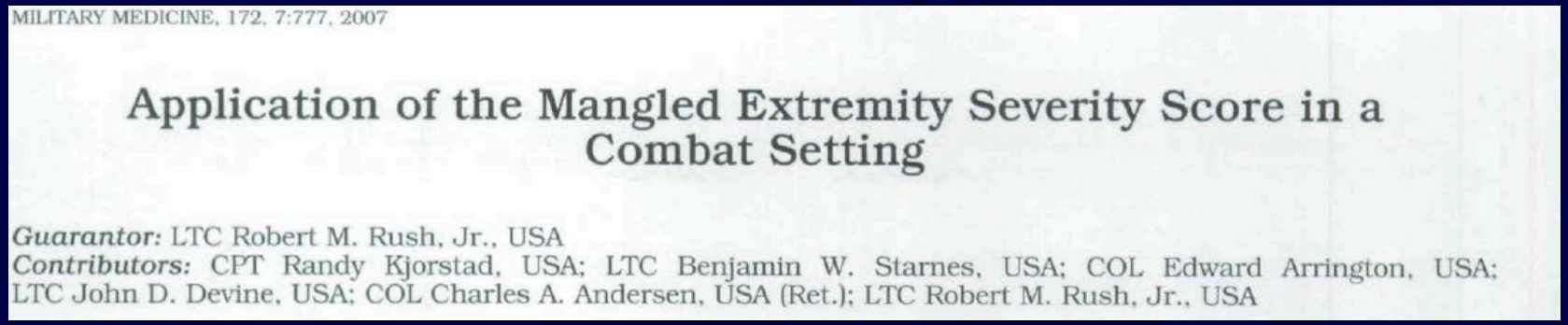


Fig. 1. Retrospective analysis of MESS values in the 250th FST during Operation

Kjorstad R, et al."Application of the Mangled Extremity Severity Score in a Mass Combat Setting." ***Military Medicine;*** 2007 Vol. 172, (7) 777-781. ©Military Medicine: International Journal of AMSUS

# Result-conclusion mismatch



MILITARY MEDICINE, 172, 7:777, 2007

Application of the Mangled Extremity Severity Score in a Combat Setting

*Guarantor:* LTC Robert M. Rush, Jr., USA
*Contributors:* CPT Randy Kjorstad, USA; LTC Benjamin W. Starnes, USA; COL Edward Arrington, USA; LTC John D. Devine, USA; COL Charles A. Andersen, USA (Ret.); LTC Robert M. Rush, Jr., USA

- Conclusion- Military MESS is helpful in determining which limbs should be amputated
- Actual results- How?
  - Look at the limbs where the uncertainty lies (those limbs scoring between 5 and 9)
  - There is statistical difference, but clinically unhelpful

# Result-conclusion mismatch

- The amputated limbs and the salvaged limbs are dissimilar groups
- Does this study show that the MESS predicts limb amputation or that the MESS is different between soldiers who got an amputation and those who did not?

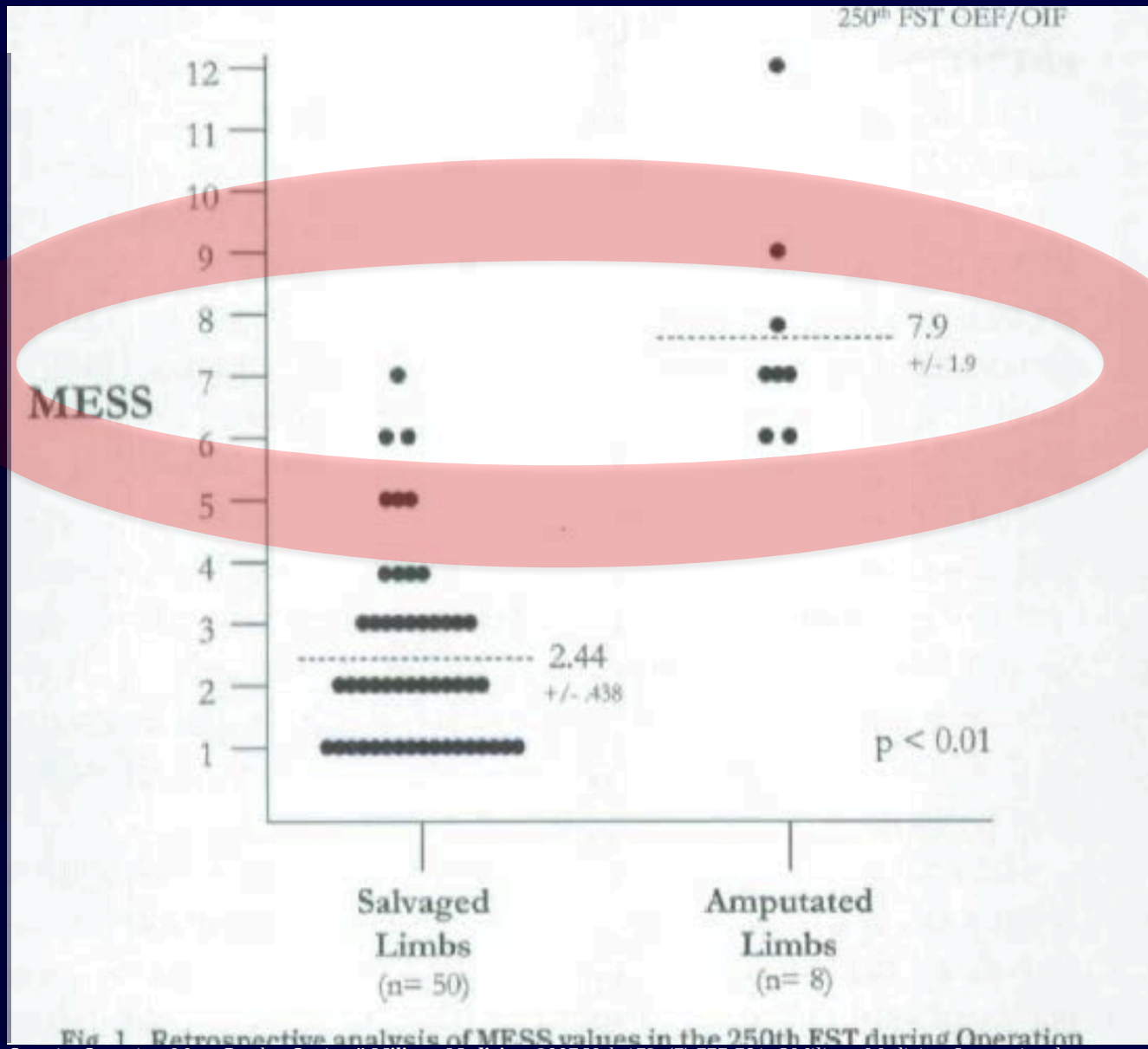



Fig. 1. Retrospective analysis of MESS values in the 250th FST during Operation

Kjorstad R, et al."Application of the Mangled Extremity Severity Score in a Mass Combat Setting." ***Military Medicine;*** 2007 Vol. 172, (7) 777-781. ©Military Medicine: International Journal of AMSUS

# Result-conclusion mismatch

**Salvaged Limbs**

- Mean
  - 2.34
- 95% Confidence Interval
  - 1.81 to 2.74
- Standard Deviation
  - 1.41

**Amputated Limbs**

- Mean
  - 7.14
- 95% Confidence Interval
  - 6.15 to 8.13
- Standard Deviation
  - 1.97

- A useful study would have to compare groups with similar means and shown different outcomes based on the MESS
- Otherwise the study is only showing that different groups of subjects with different injuries who get different treatments have different scores

# Result-conclusion mismatch

**Comparison of Manual and Gravity Stress Radiographs for the Evaluation of Supination-External Rotation Fibular Fractures**

J. Brian Gill, Timothy Risko, Viorel Raducan, J. Speight Grimes and Robert C. Schutt, Jr.
*J Bone Joint Surg Am*. 2007;89:994-999. doi:10.2106/JBJS.F.01002

- SER II
  - Nonstress radiograph
    - 3.3 +/- 0.7 (2.2 to 4.73)
  - Manual stress
    - 4.15 +/- 1.01 (2.5 to 5.67)
  - Gravity stress
    - 4.26 +/- 0.62 (3.2 to 5.25)
- SER IV
  - Nonstress radiograph
    - 3.39 +/- 0.98 (1.2 to 5)
  - Manual stress
    - 5.21 +/- 1.37 (3.2 to 7.23)
  - Gravity stress
    - 5 +/- 1.15 (3.4 to 6.6)

$p=0.55$

$p<0.02$

$p<0.05$

- There is a statistical difference between manual stress and gravity stress
  - But, is it clinically meaningful?

# Result-conclusion mismatch

**Comparison of Manual and Gravity Stress Radiographs for the Evaluation of Supination-External Rotation Fibular Fractures**

J. Brian Gill, Timothy Risko, Viorel Raducan, J. Speight Grimes and Robert C. Schutt, Jr.
*J Bone Joint Surg Am*. 2007;89:994-999. doi:10.2106/JBJS.F.01002

- SER II
  - Nonstress radiograph
    - 3.3 +/- 0.7 (2.2 to 4.73)
  - Manual stress
    - 4.15 +/- 1.01 (2.5 to 5.67)
  - Gravity stress
    - 4.26 +/- 0.62 (3.2 to 5.25)

$p=0.55$

$p<0.02$

$p<0.05$

- SER IV
  - Nonstress radiograph
    - 3.39 +/- 0.98 (1.2 to 5)
  - Manual stress
    - 5.21 +/- 1.37 (3.2 to 7.23)
  - Gravity stress
    - 5 +/- 1.15 (3.4 to 6.6)

- Look at the 95% confidence intervals of the gravity stress test
  - They are widely overlapping between the SER III and SER IV injuries
  - This suggests that the two comparison groups may not be all that different
  - Furthermore, the mean difference between groups is 0.74mm

# Result-conclusion mismatch

In conclusion, the present study demonstrates that the gravity stress radiograph is equivalent to a manual stress radiograph for determining complete deltoid ligament injury in association with an isolated distal fibular fracture and thus can be used to determine ankle stability in a patient who presents with such a fracture. ■

- Can you tell a difference of 0.74mm or less clinically?
- Study also states that gravity test is equivalent to manual test
  - No difference ≠ equivalence
  - Different methodologies and stats

# Result-conclusion mismatch

- You need to understand some statistics in order to critically evaluate papers
- ‘Someone else is the expert, I just take their word for it’
  - You are left believing whatever is written
- Many articles contain statistical errors
  - You can only find their errors if you know what to look for
  - These errors can dramatically change the perceived outcome of the study
- Multiple journals have increased their statistical reviewing processes but there is little evidence that statistical accuracy has improved

# Why do we have statistics?

- The question we want to answer is: Given these data, how likely is the null hypothesis?
- The question that a *p* value answers is: Assuming the null hypothesis is true, how unlikely are these data?
  - These two questions are different
  - We need statistics to make sure we come to the right conclusions from a study

Motulsky HJ. Common misconceptions about data analysis and statistics. *Pharmacology Research & Perspectives*. 2014.

# What is the goal of the study?

- If the study is looking to see if there is a difference between groups
  - Is one intervention/test/treatment better than another
  - Null hypothesis: no difference between groups
  - Need to determine the smallest clinically meaningful difference to power study
  - If *p* value not <.05 no difference
    - Does NOT mean the two interventions/tests/treatments are the same

Harris et al. “Not statistically different” does not necessarily mean “the same.” JBJS Am. 2012;94:e29(1-4)

# What is the goal of the study?

- If the study is trying to show two groups are equal
  - Establish that one treatment is as good as another
  - Complications, SF-36 scores, etc
  - Null hypothesis: these two treatments are different
  - Need to determine the largest difference considered clinically meaningless to power the study
  - *P* value would still need to equal <0.05 because the study would be designed to test if the treatments are different

Harris et al. JBJS Am. 2012;94:e29(1-4)

# What does *p* value mean?

- Type 1 (alpha) error: a significant association is found when there is no actually association present
- Type 2 (beta): there is no significant association found when, in reality, one exists
- *p* value refers to the alpha level. When the p value is less than 0.05, we tend to accept that a type 1 error is not being made
  - The null hypothesis is therefor rejected
- If a study shows a significant difference, one wants to make sure that the alpha level is less than 0.05

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

# The *p* value

- What is it
  - Probability test ‘alpha error’
  - *p*<.05 means 95% sure difference is true
  - May be different based on sampling bias
    - Unequally comparison groups
  - 40% of RCTs underestimated alpha error
    - Most due to not including corrections for multiple outcomes
    - 100 tests, 5% alpha error risk ➔ 5 tests ‘positive’ by chance

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

Hurwitz et al An AOA critical issue: How to read the literature to change your practice. JBJS Am 2006;88:1873-9

# The *p* value

- The *p* value gives no information about the magnitude of the association between the variables being tested
  - Only whether or not that association is likely to have occurred by chance alone
- *p* values are dichotomous, not continuous but…
  - There is likely no difference in an association of $p=0.049$ and $p=0.051$
- *p* value tells nothing about the strength of the association or the effect it may have
  - *p* value of 0.0001 shows no more effect than a *p* value of 0.049
  - A lower *p* value means the difference was less likely to occur by chance

# The *p* value

- *p* values tell of statistical significance
  - The more times a difference is searched for, the more likely a difference will be found by chance alone (increasing type 1 error)
    - This is when you need some type of correction for multiple outcome measures
- Confidence intervals can be used instead of *p* values
- Confidence intervals show many things *p* values do not
  - Statistical significance
  - Clinical significance
  - Precision of results

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

# What a *p* value is not

- If there is no difference between the groups, it does not mean that the groups are equivalent
  - You can only estimate the probability of getting certain results based on the null hypothesis being true, not vice versa
- If a study has multiple endpoints using statistical tests, a multiple comparison correction (Bonferoni) should be applied to make sure that type 1 error is not inflated
- When using small sample sizes, possibility of type 2 error increases

Strasak AM *et al* Statistical errors in medical research- a review of common pitfalls. *Swiss Med Wkly*. 2007

# Statistical mistakes relating to *p* values

- *p*-hacking
  - Running tests that were not originally designed in a hope of getting some type of significant finding
    - Adjusting the data, changing variables, etc
  - 'Floating' sample sizes
    - Increasing the sample size until a significant value is found. This skews the results because more tests would not have been run if the result was less 0.05<
  - HARKing: Generating Hypothesies After Results are Known (HARK). This leads to conflicting results because the data is used to generate the hypothesis and test it

Motulsky HJ. Common misconceptions about data analysis and statistics. *Pharmacology Research & Perspectives*. 2014.

# Statistical mistakes relating to *p* values

- *p* value has nothing to do with effect size
  - *p* value tells you there may be a difference, not how big the difference is
  - Having two means differ by 0.04 does not mean those means are any less different than if the *p* value was 0.0001
    - It only tells you how much of a chance those differences could exist to random chance

Motulsky HJ. Common misconceptions about data analysis and statistics. *Pharmacology Research & Perspectives*. 2014.

# Statistical mistakes

- There is no such thing as ‘trend towards significance’
- There have been 468 different phrases used by researchers to try to persuade the reader that the results were ‘almost’ significant
  - None of them make the differences significant

Motulsky HJ. Common misconceptions about data analysis and statistics. *Pharmacology Research & Perspectives*. 2014.

# Power

- The likelihood of finding a significant association if one truly exists
  - 1 minus the probability of type 2 (beta) error
- Most important if a study shows that a significant association does not exist
  - If the power of the study was not high enough, a true difference may actually exist
  - Typically want power to be at least 0.8
- Things that effect power
  - Sample size, effect size, variance
- About 28% of orthopaedic RCTs are underpowered
  - These may falsely reject the null hypothesis

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

Abdullah L, et al. Is There Truly “No Significant Difference”? Underpowered Randomized Controlled Trials in the Orthopaedic Literature. JBJS 2015;97:2068-73

# Statistics For Multiple Outcome Measures

- Observations must be independently calculated or have proper adjustments made for the fact that they are related
  - Otherwise, the potential for bias in either observation could be elevated
- 42% of peer reviewed studies likely had some type of bias in their statistical results by not correcting for related or multiple observations
- For example, if you are looking at patient outcomes from total knee replacements and count two knees from one patient as two separate instances of total knee replacement, the results will be biased. The outcome of the second is at least partially linked to the outcome of the first

Bryant D *et al*. How many patients? How many limbs? Analysis of patients or limbs in the orthopaedic literature: A systematic review. *J Bone J Surgery Am.* Vol 88. 2006

# Multiple Outcome measures

- When multiple endpoints are used, the *p* value should be decreased to offset the likelihood of a finding secondary to chance
- Determine a primary measure a priori and use 0.05 as the determined cutoff for that measure
- For the secondary measures
  - Most basic is a Bonferroni
    - Divide 0.05 by the number of parameters tested
    - Eg 5 secondary measures => 0.05/5=0.01 as the p value for all 5 to determine significance

Zlowodzki M and Bhandari M. Outcome Measures and Implications for Sample-Size Calculations. JBJS 2009;91 s3.35-40

# Statistical tests for articles relating to therapy

- For dichotomous variables, results can be reported as
  - Absolute risk reduction (ARR)
    - Experimental event rate (EER) minus control event rate (CER)
  - Risk difference
  - Relative Risk Reduction
    - (EER-CER)/CER
  - Hazard ratio: relative risk reduction over a period of time

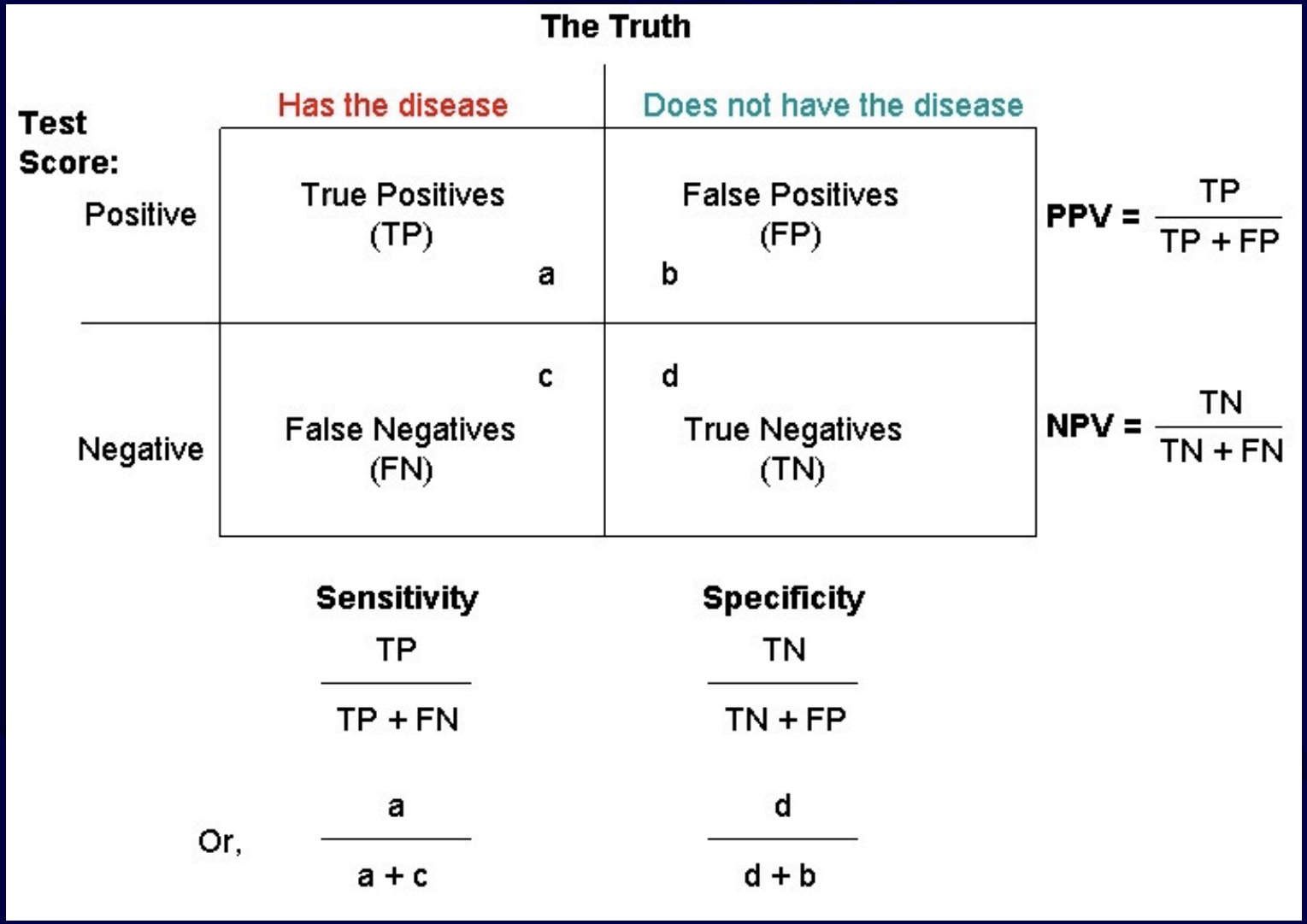
Bhandari et al. User's Guide to the Orthopaedic Literature: How to Use an Article about a Surgical Therapy. JBJS Am. 2001;83:916-26

# Statistical tests for articles relating to therapy

- For dichotomous variables, results can be reported as
  - Number Needed to Treat (NNT): 1/(relative risk difference between groups)
  - Relative risk reduction: 1-RR x 100
    - Greater relative risk, more effective therapy
  - RRR typically expressed as CI
    - CI depends on power of the study

Bhandari et al. User's Guide to the Orthopaedic Literature: How to Use an Article about a Surgical Therapy. JBJS Am. 2001;83:916-26

# Statistical tests for articles relating to diagnostic tests




http://www.med.uottawa.ca/sim/data/Sensitivity_e.htm

# Statistical tests for articles relating to diagnostic tests

- Likelihood ratio
  - For positive test= sensitivity/(1-specificity)
  - For negative test= (1-sensitivity)/specificity
  - Links the pretest probability to the posttest probability
  - Likelihood ratios of greater than 10 or less than 0.1 often have conclusive changes in posttest probability
  - Greater than 5 or less than 0.2 have moderate impact
  - Much more clinically useful than sensitivity and specificity

Bhandari et al. User's Guide to the Surgical Literature: How to Use an Article about a Diagnostic Test. JBJS Am. 2003;85:1133-40

# Loss to follow up

- The more patients that are lost to follow up, the more likely bias is introduced to the study
- A sensitivity analysis can be conducted to determine if so many patients are lost to follow up that the study is no longer valid
  - All patients lost to follow up are assumed to do poorly
  - If the results do not change, the study is valid

# When patients are lost to follow up

- Losing patients to follow up can bias a study's results
- Three ways to analyze data when patients were lost to follow up
  - Intention to treat analysis
  - Per-protocol analysis
  - Treatment-received analysis

# Intention to treat

- Groups are analyzed in regards to their allocated group regardless of whether or not they completed their prescribed treatment
  - Preserves randomization
  - Minimizes type 1 error
  - Makes most conservative estimate of treatment effect and may increase type 2 error

- Excluded patients often have worse prognosis
  - Too sick to get the operation they were assigned to get➔ go to ‘control’ group ➔ invalid picture of an operation that ‘works’

Bubbar et al. The Intention-to-Treat Principle: A Primer for the Orthopaedic Surgeon. JBJS Am. 2006;88:2097-99

# Per-protocol analysis

- Excluding any subjects from data analysis that violated the study protocol (crossovers, lost to follow up, etc)
  - This may leave the residual groups that are analyzed as dissimilar
  - It undermines randomization and may introduce bias
  - It may also cause the treatment effect to be over-estimated

Bubbar et al. The Intention-to-Treat Principle: A Primer for the Orthopaedic Surgeon. JBJS Am. 2006;88:2097-99

# Treatment-Received analysis

- Subjects are evaluated based on the treatment they receive, not what they were assigned
  - Similar to per-protocol but instead of excluding them altogether they are analyzed

Bubbar et al. The Intention-to-Treat Principle: A Primer for the Orthopaedic Surgeon. JBJS Am. 2006;88:2097-99

# In Summary

- When determining what to read: does it apply to my current practice or is it for future knowledge?
- When determining how to read: identify your reading goals before reading
- Determine what the article is trying to tell you, than analyze the article critically
- Learning basic statistics will allow you to determine if an article's conclusions match its results or if there is a mismatch