CRITICAL APPRAISAL
OF THE SCIENTIFIC LITERATURE:
Sorting Out the Good, the Bad, and the Ugly

Scott E. Moser, MD
Associate Professor
Department of Family and Community Medicine
The University of Kansas School of Medicine-Wichita
(316) 293-2607
smoser@kumc.edu
WHAT YOU SHOULD KNOW AFTER THIS PRESENTATION:

Critical Appraisal of the Scientific Literature

You should be able to answer the following questions on scientific articles:

1. Did the Introduction/Background provide:
   a. A logical and convincing rationale
   b. The statement of the problem
   c. Purpose, questions, or hypotheses that are a logical extension of the rationale

2. Methods – were the following appropriate:
   a. Study design
   b. Subject selection
   c. Methods of measurement
   d. Analytical Techniques

3. Results: What is the clinical and statistical significance?

4. Discussion/Conclusions:
   a. Did the discussion relate to the problem?
   b. Were the conclusions drawn fairly?
   c. Was the explanation of the significance reasonable?
   d. Were there objective comparisons to previous research?
   e. Did the authors discuss the limitations of the study?
   f. Did the authors make implications for practice?
   g. Did the authors make suggestions for future research?

Other

5. How should this study change my practice?
6. Who can help you with appraisal of the literature?
Critical Appraisal of the Scientific Literature:

Sorting Out the Good, the Bad, and the Ugly

Scott Moser, MD

Objective:

- Review and critically evaluate scientific literature related to initiating a research project.

Overview:

- Compare/contrast usual EBM to the special case of preparing your own study
- Focus on the most common problems in doing critical appraisal
- Give you resources to apply later
- Practical examples

Before You Begin:

- Ask, “Why am I going there?”
- Assemble your tool kit.

Why go there?

- Has my question already been answered?
- Can I add to what has already been done?

Why go there?

- Did someone steal my idea before I even had it?
- If so, did they do a crummy enough job that I can do better?
Why go there?

• Methodology ideas: How will I study my question?
• Introduction ideas: What is known? Not known? Framework for my question
• Discussion ideas: How do my results compare/contrast with others?

Toolkit Essentials:

• Reference text on critical appraisal (not a statistics text)
• Reference person

General EBM Approach:

• Are the study results valid?
• Are the valid results important?
• Does this valid, important evidence apply to my case?

Question 1:

Why was the study done and what hypotheses were the authors testing?

Question 2:

What type of study was done?

Primary Studies:

• Experiments
• Clinical trials
• Surveys
Secondary Research:
- Overviews
  - Non-systematic review
  - Systematic review
  - Meta-analysis
- Guidelines
- Decision analyses
- Economic analyses

Question 3:
Was this design appropriate to the broad field of research studied?

Broad Topics of Research:
- Therapy
- Diagnosis
- Screening
- Prognosis
- Causation/Harm

Hierarchy of Evidence:
- Systematic review/meta-analysis
- Randomized controlled trial with definitive results
- Randomized controlled trial with non-definitive results
- Cohort study
- Case-control study
- Cross sectional survey
- Case series/Case report

Preferred Study Design:
- Therapy: RCT>cohort>case control>case series
- Diagnosis: Prospective, blind comparison to a gold standard
- Screening/Prevention: RCT>cohort>case control>case series
- Prognosis: Cohort>case control>case series
- Causation/Harm: RCT>cohort>case control>case series (depending on disease rarity, case reports may be crucial)
- Clinical Exam: Prospective, blind comparison to gold standard
- Cost: Economic analysis

Methods:
- Who was the study about?
- Was the design of the study sensible?
- Was systematic bias avoided or minimized?
- Was assessment “blind”? 
Statistics for the Rest of Us:

• Sample size large enough?
• Duration of follow-up long enough?
• Follow-up complete enough? (Too many dropouts?)

Additional Questions:

• Results: Statistical vs. clinical significance
• Discussion: Did the authors overstate or understate their conclusions?
• General: So what? How should this study change my practice? How does this study change my study?

Keys to Diagnosis Articles:

• Clear comparison groups, all suspected with disorder, some free of it.
• “Gold standard” applied to all participants.
• Blinded assessment of test and gold standard.
• Likelihood ratios are more useful than sensitivity and specificity (Bayes’ theorem).

Likelihood Ratios:

• >10 or <0.1        Cause large changes
• 5-10 or 0.1-0.2    Moderate changes
• 2-5 or 0.2-0.5     Small changes
• <2 or >0.5        Tiny changes
• 1.0                No change at all

Keys to Therapy Articles:

• Random allocation to comparison groups.
• Outcome measure of known or probable clinical importance.
• Follow-up of > 80% of subjects.
• Consider Number Needed to Treat and side-effect profile.

Keys to Prognosis Articles:

• Inception cohort: early in the course of the disorder and free of the outcome.
• Objective assessment of clinically important outcomes.
• Follow-up of > 80% of subjects.
### Keys to Causation/Harm Articles:
- Clearly identified comparison group for those at risk for or those having the outcome.
- Blinding of observers of outcome to exposure and vice versa.

### Keys to Review Articles:
- Comprehensive search for relevant articles.
- Explicit criteria for rating relevance and merit of studies.
- Inclusion of all relevant studies.

### Examples:
Appendix A: Checklists for finding, appraising, and implementing evidence

Unless otherwise stated, these checklists can be applied to randomised controlled trials, other controlled clinical trials, cohort studies, case-control studies, or any other research evidence.

Is my practice evidence based?—a context-sensitive checklist for individual clinical encounters (see chapter 1)

1 Have I identified and prioritised the clinical, psychological, social, and other problem(s), taking into account the patient’s perspective?

2 Have I performed a sufficiently competent and complete examination to establish the likelihood of competing diagnoses?

3 Have I considered additional problems and risk factors that may need opportunistic attention?

4 Have I, when necessary, sought evidence (from systematic reviews, guidelines, clinical trials, and other sources) pertaining to the problems?

5 Have I assessed and taken into account the completeness, quality, and strength of the evidence?

6 Have I applied valid and relevant evidence to this particular set of problems in a way that is both scientifically justified and intuitively sensible?

7 Have I presented the pros and cons of different options to the patient in a way they can understand and incorporated the patient’s utilities into the final recommendation?

8 Have I arranged review, recall, referral, or other further care as necessary?
Checklist for searching Medline or the Cochrane library (see chapter 2)

1. To look for an article you know exists, search by textwords (in title, abstract, or both) or use field suffixes for author, title, institution, journal, and publication year.

2. For a maximally sensitive search on a subject, search under both MeSH headings (exploded) and textwords (title and abstract), then combine the two by using the Boolean operator “or”.

3. For a focused (specific) search on a clear cut topic, perform two or more sensitive searches as in (2), and combine them by using the Boolean operator “and”.

4. To find articles that are likely to be of high methodological quality, insert an evidence based quality filter for therapeutic interventions, aetiology, diagnostic procedures, or epidemiology (see appendix B) and/or use maximally sensitive search strategies for randomised trials, systematic reviews, and meta-analyses (see appendix C).

5. Refine your search as you go along—for example, to exclude irrelevant material, use the Boolean operator “not”.

6. Use subheadings only when this is the only practicable way of limiting your search as manual indexers are fallible and misclassifications are common.

7. When limiting a large set, browse through the last 50 or so abstracts yourself rather than expecting the software to pick the best half dozen.

Checklist to determine what a paper is about (see chapter 3)

1. Why was the study done (what clinical question did it examine)?

2. What type of study was done?

Checklist for the methods section of a paper (see chapter 4)

1. Was the study original?

2. Whom is the study about?
   - How were subjects recruited?
   - Who was included in and who was excluded from the study?
   - Were the subjects studied in “real life” circumstances?

3. Was the design of the study sensible?
   - What intervention or other manoeuvre was being considered?
   - What outcome(s) were measured and how?

4. Was the study adequately controlled?
   - If a “randomised trial” was randomisation truly random?
   - If a cohort, case-control, or other non-randomised comparative study were the controls appropriate?
   - Were the groups comparable in all important aspects except for the variable being studied?
   - Was assessment of outcome (or, in a case-control study, allocation of caseness) “blind”?
HOW TO READ A PAPER

5. Was the study large enough and continued for long enough, and was follow up complete enough, to make the results credible?

Checklist for the statistical aspects of a paper (see chapter 5)

1. Have the authors set the scene correctly?
   - Have they determined whether their groups are comparable and, if necessary, adjusted for baseline differences?
   - What sort of data have they got and have they used appropriate statistical tests?
   - If the statistical tests in the paper are obscure why have the authors chosen to use them?
   - Have the data been analysed according to the original study protocol?

2. Paired data, tails, and outliers:
   - Were paired tests performed on paired data?
   - Was a two tailed test performed whenever the effect of an intervention could conceivably be a negative one?
   - Were outliers analysed with both common sense and appropriate statistical adjustments?

3. Correlation, regression and causation:
   - Has correlation been distinguished from regression and has the correlation coefficient (r value) been calculated and interpreted correctly?
   - Have assumptions been made about the nature and direction of causality?

4. Probability and confidence:
   - Have P values been calculated and interpreted appropriately?
   - Have confidence intervals been calculated and do the authors’ conclusions reflect them?

5. Have the authors expressed their results in terms of the likely harm or benefit that an individual patient can expect, such as:
   - Relative risk reduction
   - Absolute risk reduction
   - Number needed to treat
   - Odds ratio.

Checklist for material provided by a pharmaceutical company representative (see chapter 6)

1. Does this material cover a subject that interests me and is clinically important in my practice?

2. Has this material been published in independent peer reviewed journals? Has any significant evidence been omitted from this presentation or withheld from publication?

3. Does the material include high level evidence such as systematic reviews, meta-analyses, or double blind randomised controlled trials against the drug’s closest competitor given at optimal dosage?

4. Have the trials or reviews examined a clearly focused, important and answerable clinical question that reflects a problem of relevance to patients? Do they provide evidence on safety, tolerability, efficacy, and price?

5. Has each trial or meta-analysis defined the condition to be treated, the patients to be included, the interventions to be compared, and the outcomes to be examined?

6. Does the material provide direct evidence that the drug will help my patients live a longer, healthier, more productive, or symptom free life?

7. If a surrogate outcome measure has been used, what is the evidence that it is reliable, reproducible, sensitive, specific, a true predictor of disease, and rapidly reflects the response to therapy?
8 Do trial results indicate whether (and how) the effectiveness of the treatments differed and whether there was a difference in the type or incidence of adverse reactions? Are the results expressed in terms of numbers needed to treat, and are they clinically as well as statistically significant?

9 If large amounts of material have been provided by the representative, which three papers provide the strongest evidence for the company’s claims?

Checklist for a paper that claims to validate a diagnostic or screening test (see chapter 7)

1 Is this test potentially relevant to my practice?
2 Has the test been compared with a true gold standard?
3 Did this validation study include an appropriate spectrum of subjects?
4 Has work up bias been avoided?
5 Has observer bias been avoided?
6 Was the test shown to be reproducible both within and between observers?
7 What are the features of the test as derived from this validation study?
8 Were confidence intervals given for sensitivity, specificity, and other features of the test?
9 Has a sensible “normal range” been derived from these results?
10 Has this test been placed in the context of other potential tests in the diagnostic sequence for the condition?

Checklist for a systematic review or meta-analysis (see chapter 8)

1 Did the review examine an important clinical question?
2 Was a thorough search done of the appropriate database(s) and were other potentially important sources explored?

3 Was methodological quality assessed and the trials weighted accordingly?
4 How sensitive are the results to the way the review has been done?
5 Have the numerical results been interpreted with common sense and due regard to the broader aspects of the problem?

Checklist for a set of clinical guidelines (see chapter 9)

1 Did the preparation and publication of these guidelines involve a significant conflict of interest?
2 Are the guidelines concerned with an appropriate topic, and do they state clearly the goal of ideal treatment in terms of health and/or cost outcome?
3 Was the guideline development panel headed by a leading expert in the field (ideally it should not be) and was a specialist in the methodology of secondary research (for example, meta-analyst, health economist) involved?
4 Have all the relevant data been scrutinised and do the guidelines’ conclusions seem to be in keeping with the data?
5 Do they cover variations in medical practice and other controversial areas (for example, optimum care in response to genuine or perceived underfunding)?
6 Are the guidelines valid and reliable?
7 Are they clinically relevant, comprehensive, and flexible?
8 Do they take into account what is acceptable to, affordable by, and practically possible for patients?
9 Do they include recommendations for their own dissemination, implementation, and periodic review?

Checklist for an economic analysis (see chapter 10)

1 Is the analysis based on a study that answers a clearly defined clinical question about an economically important issue?
2 Whose viewpoint are costs and benefits considered from?
HOW TO READ A PAPER
3 Have the interventions being compared been shown to be clinically effective?
4 Are the interventions sensible and workable in the settings where they are likely to be applied?
5 Which method of economic analysis was used and was this appropriate?
   • If the interventions produced identical outcomes ⇒ cost-minimisation analysis
   • If the important outcome is unidimensional ⇒ cost-effectiveness analysis
   • If the important outcome is multidimensional ⇒ cost-utility analysis
   • If the cost-benefit equation for this condition needs to be compared with cost-benefit equations for different conditions ⇒ cost-benefit analysis
   • If a cost-benefit analysis would otherwise be appropriate but the preference values given to different health states are disputed or likely to change ⇒ cost-consequences analysis.
6 How were costs and benefits measured?
7 Were incremental rather than absolute benefits compared?
8 Was health status in the “here and now” given precedence over health status in the distant future?
9 Was sensitivity analysis performed?
10 Were “bottom line” aggregate scores overused?

Checklist for a qualitative research paper (see chapter 11)
1 Did the article describe an important clinical problem examined via a clearly formulated question?
2 Was the qualitative approach appropriate?
3 How were the setting and the subjects selected?

Checklist for health care organisations working towards an evidence based culture for clinical and purchasing decisions (see chapter 12 and reference 38 of that chapter)

1 Leadership—How often has information on effectiveness or evidence based medicine been discussed at board meetings in the past 12 months? Has the board taken time out to learn about developments in clinical and cost effectiveness?
2 Investment—What resources are the organisation investing in finding and using clinical effectiveness information? Is there a planned approach to promoting evidence based medicine that is properly resourced and staffed?
3 Using available resources—What action has been taken by the organisation in response to EL(93)115 (Improving Clinical Effectiveness) and EL(94)74 (Improving the Effectiveness of the NHS)? What has changed in the organisation as a result?
4 Implementation—Who is responsible for receiving, acting on, and monitoring the implementation of Effective Health Care bulletins? What action has been taken on each of the bulletins issued to date?
HOW TO READ A PAPER

5 Clinical guidelines—Who is responsible for receiving, acting on, and monitoring clinical practice guidelines? Do those arrangements ensure that both managers and clinicians play their part in guideline development and implementation?

6 Training—Has any training been provided to staff within the organisation (both clinical and non-clinical) on appraising and using evidence of effectiveness to influence clinical practice?

7 Contracts—How often does clinical and cost-effectiveness information form an important part of contract negotiation and agreement? How many contracts contain terms that set out how effectiveness information is to be used?

8 Incentives—What incentives—both individual and organisational—exist to encourage the practice of evidence based medicine? What disincentives exist to discourage inappropriate practice and unjustified variations in clinical decision making?

9 Information systems—Is the potential of existing information systems to monitor clinical effectiveness being used to the full? Is there a business case for new information systems to deal with the task, and is this issue being considered when purchasing decisions about information technology are made?

10 Clinical audit—Is there an effective clinical audit programme throughout the organisation, capable of examining issues of clinical effectiveness and bringing about appropriate changes in practice?
Evidence Based Medicine
Glossary

2X2 Table

<table>
<thead>
<tr>
<th></th>
<th>+ Disease</th>
<th>- Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Exposure</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>- Exposure</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>a + c</td>
<td>b + d</td>
<td></td>
</tr>
</tbody>
</table>

**Absolute risk increase**: The increase in risk of disease expressed by the absolute arithmetic difference between the Experimental Event Rate and the Control Event Rate ($\frac{a}{a+b} - \frac{c}{c+d}$).

**Absolute risk reduction**: The reduction in risk of disease expressed by the absolute arithmetic difference between the Control Event Rate and the Experimental Event Rate ($\frac{c}{c+d} - \frac{a}{a+b}$).

**Association**: The occurrence together of two or more characteristics or events more often than would be expected by chance.

**Blind assignment**: Blind assignment occurs when individuals are assigned to a study group and a control group with neither the investigator nor the subjects being aware of the group to which they are assigned. When both investigator and subjects are “blinded,” the study is sometimes referred to as a double-blind study.

**Case control study (or Retrospective study)**: A study that begins by identifying cases and controls and compares them to determine the proportion of each group that previously possessed a characteristic or an exposure to a factor.

**Cohort**: A group of individuals who share a common exposure or experience. Applies to prospective studies in which the study group and the control group each share a common experience. Also applies to a cohort effect in which changes in rates are explained by a unique experience shared by a subgroup of the population.

**Cohort study (or Prospective study)**: A study in which a group classified by the presence or absence of a factor is followed to determine whether those with the factor have a greater or lesser likelihood of developing the disease or condition under study. Prospective studies can be done concurrently or nonconcurrently.

**Confidence interval**: A range of values within which the true value lies for the whole population from whom the study subjects were selected. The confidence interval narrows as the number of patients on which it is based increases.

*See 2x2 Table*
**Con Founding variable**: A variable that differs between the study group and the control group, affecting the outcome being assessed.

**Control Event Rate (CER)**: The proportion of persons who actually have the outcome among all who have not been exposed to a treatment or risk factor. \( \frac{c}{c+d} \).

**Control group**: A population or sample of a population that is used for comparison with a study group. Ideally, the control group is identical to the study group except that it does not possess the characteristic or has not been exposed to the treatment under study.

**Cross-sectional study**: A study that begins by identifying individuals with and without the condition or disease under study. It then compares them to determine the proportion of each group that currently possesses a characteristic or is currently being exposed to a factor.

**Experimental Event Rate (EER)**: The proportion of persons who actually have the outcome among all those exposed to the treatment or risk factor. \( \frac{a}{a+b} \).

**Experimental study (or clinical trial)**: A study in which the investigator assigns individuals to both a study and a control group prior to the occurrence of the study condition or outcome. Only experimental studies allow random and blind assignment of individuals to study and control groups.

**False negative**: When an individual’s value on a test is negative but the disease or condition is present as determined by the gold standard, then the test has produced a false negative \( \frac{c}{c} \).

**False positive**: When an individual’s value on a test is positive but the disease or condition is absent as determined by the gold standard, then the test has produced a false positive \( \frac{b}{b} \).

**Gold standard**: The criteria used to unequivocally define the presence of a condition or disease under study.

**Incidence rate**: An approximation of the risk of developing a given disease. Calculated by taking the ratio of the number of individuals who develop the condition over a period of time, divided by the number of individuals in the population at the midpoint of the time interval.

**Intention-to-treat**: Analyzing all participants in their original randomization, even if they withdrew, or crossed over from treatment to no treatment.

**Likelihood ratio for a positive test result**: The likelihood that a person with a disease will have a positive test result compared to the likelihood that a person without a disease will have a positive test result. \( \text{sensitivity}/(1-\text{specificity}) \).

**Likelihood ratio for a negative test result**: The likelihood that a person with a disease will have a negative test result compared to the likelihood that a person without a disease will have a negative test result. \( (1-\text{sensitivity})/\text{specificity} \).

*See 2x2 Table*
**Matching**: A procedure used at the beginning of studies that selects control individuals who possess the same characteristic as study individuals. Matching is done to control for a confounding variable and to reduce the number of individuals required for a study.

**Number needed to harm (NNH)**: The number of people who must receive the exposure to create one additional adverse outcome in comparison with no exposure ($\frac{1}{a/(a+b)} - \frac{c/(c+d)}{a/(a+b)}$) or $1/$Absolute Risk Increase

**Number needed to treat (NNT)**: The number of people who must receive the exposure to create one additional improved outcome in comparison with no exposure ($\frac{1}{c/(c+d)} - \frac{a/(a+b)}{a/(a+b)}$) or $1/$Absolute Risk Reduction

**Odds ratio**: The odds ratio is a measure of the degree or strength of an association applicable to all types of studies employing nominal data, but usually applied to retrospective and cross-sectional studies. The odds ratio for retrospective and cross-sectional studies is measured as the odds of having the risk factor if the condition is present, divided by the odds of having the risk factor if the condition is not present. ($\frac{ad}{bc}$).

**p value**: The probability of making a Type I or alpha error, to reject the null hypothesis when it is true. "$p < .05$" means that: (1) the experimental hypothesis has been accepted as true and the results between groups are significantly different and (2) the probability of an error that the null hypothesis is true is less than 5%.

**Post-test odds**: The odds of a diseased person having a positive (or negative) test result based upon the pre-test odds and the likelihood ratio. ($\text{pre-test odds} \times \text{likelihood ratio}$).

**Post-test probability**: Probability of disease in a person given that a positive test occurs. ($\frac{\text{post-test odds}}{1+\text{post-test odds}}$).

**Power of statistical significance test**: The ability of the test to detect a statistically significant difference or association in sample data if a true difference or association is present in the larger parent populations.

**Predictive value of a negative test**: The proportion of those with a negative test who actually do not have a condition or disease as measured by the gold standard. This measure incorporates the prevalence of the condition or disease. Clinically, the predictive value of a negative test is the probability that an individual is free of the disease if the test is negative. ($\frac{d}{c+d}$).

**Predictive value of a positive test**: The proportion of those with a positive test who actually have the condition or disease as measured by the gold standard. This measure incorporates the prevalence of the condition or disease. Clinically, the predictive value of a positive test is the probability that an individual has the disease if the test is positive. ($\frac{a}{a+b}$).

**Pre-test odds**: The odds of a diseased person having a positive test result based upon the pre-test probability or prevalence. ($\frac{a+c}{b+d}$).

**Pre-test probability**: Prevalence or the proportion of persons in a defined population at a given point in time with the disease in question. ($\frac{a+c}{a+b+c+d}$).

*See 2x2 Table
**Prevalence:** An approximation of the risk of having a given disease at a given point in time. The prevalence rate is calculated by taking the ratio of those who have the condition at a point in time divided by the number of individuals in the population at the same point in time. As used in diagnostic tests, the prevalence refers to the probability of the disease being present before the test is performed. Prevalence in this sense also is called pretest probability or prior probability. \( \frac{a+c}{a+b+c+d} \).*

**Random assignment (or randomization):** Individuals chosen for the study have equal chance of being assigned either to the study group or to the control group. Random assignment is only possible in an experimental study. Randomization does not assure that the study group or the control group will be identical with respect to all characteristics that affect the outcome of an investigation.

**Random sample:** A sample in which each individual or entity in the larger population of interest has an equal opportunity of being included in the sample. Statistical significance testing requires that the study group and the control group each be random samples of all those who could have been selected.

**Relative risk:** The probability of developing the outcome if the risk factor is present divided by the probability of developing the outcome if the risk factor is not present. The relative risk is a measure of the strength or degree of association applicable to prospective and experimental studies. \( \frac{a/(a+b)}{c/(c+d)} \).*

**Relative risk increase:** The proportional increase in rates of disease between persons exposed to a risk factor and those not exposed \( \frac{a/(a+b)-c/(c+d)}{c/(c+d)} \).*

**Relative risk reduction:** The proportional reduction in rates of disease between persons exposed to a treatment and those not exposed \( 1-\frac{a/(a+b)}{c/(c+d)} \).*

**Risk of occurrence:** Probability of developing a given disease or condition as measured by the number of cases occurring over a period of time divided by the number of individuals in the population during the same period of time. This measure is known as absolute risk in contrast to relative risk that compares the risk to those exposed to or possessing a study factor versus those who do not. \( \frac{a}{a+b} \).*

**Selection bias:** An error in assignment that occurs whenever the selection of study and control individuals produces a difference between the groups that affects the results of the study.

**Sensitivity:** The proportion of those with the disease or condition as measured by the gold standard who are positive by the test being studied. \( \frac{a}{a+c} \).*

**Specificity:** The proportion of those without the disease or condition as measured by the gold standard who are negative by the test being studied. \( \frac{d}{b+d} \).*

**Statistical significance test:** Statistical technique for determining the probability that the observed or larger differences or associations would occur by chance alone if there is no true difference or association in the larger parent population.

**True negative:** A true negative indicates that the individual did not have the disease or condition as measured by either the gold standard or the test being studied \( \frac{d}{d} \).*

*See 2x2 Table
**True positive**: A true positive indicates that the individual has the disease or condition as measured by either the gold standard or the test being studied (*αf)*.

**Type I error**: An error that occurs when using data from samples that demonstrates a statistically significant difference or association when no true difference or association exists in the parent populations (alpha *α* error).

**Type II error**: An error that occurs from failure to demonstrate a statistically significant difference or association in sample data when one actually exists in the larger parent populations (beta *β* error).

*See 2x2 Table*
“Critical Appraisal of the Scientific Literature: Sorting Out the Good, the Bad, and the Ugly”

References:


JAMA “Users’ Guides to the Medical Literature” series:

I. How to Get Started. 270(17):2093-5.
II. How to Use an Article About Therapy or Prevention
   A. Are the Results of the Study Valid? 270(21):2598-2601; Dec 1, 1993.
III. How to Use an Article About a Diagnostic Test

(These can be accessed at: http://www.cche.net/usersguides/main.asp)
The Valenti procedure for hallux limitus: a long-term follow-up and analysis.
Kurtz DH, Harrill JC, Kaczander BI, Solomon MG.
Botsford General Hospital, Farmington Hills, MI, USA.

A retrospective analysis of the long-term efficacy of the Valenti procedure for hallux limitus was performed from 1989 to 1997. A total of 33 patients (36 procedures) were selected, surveyed, and examined. Preoperative and postoperative radiographic evaluations and levels of function and pain were obtained from medical records. Complications and patient satisfaction data were collected and reviewed. The patients were evaluated clinically and radiographically. The average age of the patient at the time of surgery was 50.6 years (range, 35-75 years) with an average follow-up of 4.16 years (range, 1-9 years). The average grade of hallux limitus/rigidus was grade II. Five patients were classified as grade I, 23 as grade II, five as grade III, and none as grade IV, based on the modified Drago/Regnauld grading system. Subjective results were calculated based on the American Orthopedic Foot and Ankle Society clinical rating system. Twenty-two patients were rated as having an excellent result, 11 had good results, two had fair results, and one patient had a poor result. We conclude that the Valenti arthroplasty for symptomatic hallux rigidus/limitus is a good procedure for arthritic and degenerated first metatarsophalangeal joints where implant arthroplasty, osteotomy, or arthrodesis are not viable options. Advantages include increased range of motion, decreased postoperative pain, technical ease of performance, maintenance of intrinsic musculature, and rapid return to closed footgear.
Hallux Rigidus Grading and Long-Term Results of Operative Treatment

Michael J. Coughlin, MD1 and Paul S. Shurnas, MD2

1 901 North Curtis Road, Suite 503, Boise, ID 83706. E-mail address: footmd@aol.com
2 Regional Orthopaedic Health Care, No. 3 Medical Plaza, Mountain Home, AR 72653

Investigation performed at St. Alphonsus Regional Medical Center, Boise, Idaho

Background: There have been few long-term studies documenting the outcome of surgical treatment of hallux rigidus. The purposes of this report were to evaluate the long-term results of the operative treatment of hallux rigidus over a nineteen-year period in one surgeon's practice and to assess a clinical grading system for use in the treatment of hallux rigidus.

Methods: All patients in whom degenerative hallux rigidus had been treated with cheilectomy or metatarsophalangeal joint arthrodesis between 1981 and 1999 and who were alive at the time of this review were identified and invited to return for a follow-up evaluation. At this follow-up evaluation, the hallux rigidus was graded with a new five-grade clinical and radiographic system. Outcomes were assessed by comparison of preoperative and postoperative pain and AOFAS (American Orthopaedic Foot and Ankle Society) scores and ranges of motion. These outcomes were then correlated with the preoperative grade and the radiographic appearance at the time of follow-up.

Results: One hundred and ten of 114 patients with a diagnosis of hallux rigidus returned for the final evaluation. Eighty patients (ninety-three feet) had undergone a cheilectomy, and thirty patients (thirty-four feet) had had an arthrodesis. The mean duration of follow-up was 9.6 years after the cheilectomies and 6.7 years after the arthrodeses. There was significant improvement in dorsiflexion and total motion following the cheilectomies (p = 0.0001) and significant improvement in postoperative pain and AOFAS scores in both treatment groups (p = 0.0001). A good or excellent outcome based on patient self-assessment, the pain score, and the AOFAS score did not correlate with the radiographic appearance of the joint at the time of final follow-up. Dorsiflexion stress radiographs demonstrated correction of the elevation of the first ray to nearly zero. There was no association between hallux rigidus and hypermobility of the first ray, functional hallux limitus, or metatarsus primus elevatus.

Conclusions: Ninety-seven percent (107) of the 110 patients had a good or excellent subjective result, and 92% (eighty-six) of the ninety-three cheilectomy procedures were successful in terms of pain relief and function. Cheilectomy was used with predictable success to treat Grade-1 and 2 and selected Grade-3 cases. Patients with Grade-4 hallux rigidus or Grade-3 hallux rigidus with <50% of the metatarsal head cartilage remaining at the time of surgery should be treated with arthrodesis.

Level of Evidence: Therapeutic study, Level IV (case series [no, or historical, control group]). See Instructions to Authors for a complete description of levels of evidence.

The surgical options for hallux rigidus in the presence of painful but moderate degenerative metatarsophalangeal joint disease are limited to either joint-destructive or joint-preserving procedures. The following study compared the effectiveness of 2 joint-preservation procedures. Forty-nine patients, with a mean age of 53 years, underwent phalangeal osteotomy and were reviewed at an average 29 months postoperatively. A subsequent group of 59 patients, with a mean age of 51 years, underwent first metatarsal decompression osteotomy and were reviewed at an average 15 months postoperatively. In the phalangeal osteotomy group, 65% of patients were completely satisfied, 24% were satisfied with reservation, and 11% were dissatisfied. Three patients suffered continued metatarsophalangeal joint pain, 3 developed hallux interphalangeal joint pain, and 4 patients developed transfer metatarsalgia. The postoperative decrease from 36 degrees to 35 degrees in mean peak hallux dorsiflexion on walking was not significant. In the first metatarsal decompression osteotomy group, 54% were completely satisfied, 13.5% were satisfied with reservations, and 32% were dissatisfied. Continued metatarsophalangeal joint pain occurred in 2 patients, 18 developed transfer metatarsalgia, and 6 of these patients required lesser metatarsal osteotomy. Peak hallux dorsiflexion during walking increased from 36 degrees to 42 degrees (P < .001). First metatarsal decompression osteotomy will increase joint range of motion but the risk of complication and patient dissatisfaction is less after phalangeal osteotomy. Neither procedure could be considered definitive for hallux rigidus.

Publication Types:
Comparative Study
The Frykman, Melone, Mayo, and AO classification systems for distal radius fractures were evaluated for interobserver reliability and intraobserver reproducibility in a clinical setting using initial plain radiographs. Two attending orthopedic hand surgeons and two attending radiologists classified 55 sets of distal radius fractures. Kappa-statistics were used to establish a relative level of agreement between observers for the two readings and between separate readings by the same observer. Interobserver agreement was rated as moderate for the Mayo classification and fair for the Frykman, Melone, and AO classifications. Intraobserver agreement was substantial for only one of four observers for each of the Frykman, Melone, and Mayo, while the remaining three observers achieved only fair to moderate reproducibility. Intraobserver agreement for the AO classification was fair for all four of the observers. Neither interobserver or intraobserver agreement was affected by combining similar subclasses in the Melone classification or by reducing the number of categories in the AO system from 27 to 9. However, further reducing the AO system to its three main types brought agreement to the "substantial" level. No difference was found in interobserver agreement between the first and second readings or in interobserver or intraobserver agreement between orthopedic hand surgeons and radiologists. Understanding the limitations of fracture classifications based solely on plain radiographs can help avoid undue reliance on them. Given the low degree of interobserver and intraobserver agreement for each of the distal radius fracture classifications in this study, their use as the sole means for determining the direction of treatment or for the direct comparison of results among different studies is not warranted.
Current Concepts Review
Hallux rigidus and osteoarthrosis of the first metatarsophalangeal joint.
Shereff MJ, Baumhauer JF.
Department of Orthopaedics, Medical College of Wisconsin, Milwaukee 53226, USA.

No abstract