Assessing risk of bias in clinical research studies of interventions
Roger Chou MD

Examining the reliability of published research findings

Roger Chou, MD
Associate Professor of Medicine
Department of Medicine and Department of Medical Informatics
and Clinical Epidemiology
Scientific Director, the Oregon Evidence-based Practice Center

Learning objectives

- Discuss sources of bias in research studies and how they affect results and conclusions
- Describe how to critically appraise research studies and determine when findings can be applied to clinical practice
- Learn when to trust and use systematic reviews and guidelines

Mathematically modeled the likelihood that a “positive” research finding is actually true

Based on Bayes’ theorem for updating probabilities
- The post-test probability of a true finding depends on
  - The pre-test probability
  - The statistical power of the study
  - The extent of bias in the study

Estimated probabilities of true positive findings

<table>
<thead>
<tr>
<th>Power</th>
<th>Pretest probability</th>
<th>Bias</th>
<th>Example</th>
<th>Probability that the finding is true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.20</td>
<td>Low</td>
<td>Discovery-oriented exploratory research with limited bias</td>
<td>0.0015</td>
</tr>
<tr>
<td>High</td>
<td>0.80</td>
<td>Low</td>
<td>Adequately powered exploratory epidemiological study</td>
<td>0.20</td>
</tr>
<tr>
<td>Low</td>
<td>0.20</td>
<td>Moderate</td>
<td>Underpowered, poorly performed RCT</td>
<td>0.17</td>
</tr>
<tr>
<td>Low</td>
<td>0.20</td>
<td>Moderate</td>
<td>Underpowered, well performed RCT</td>
<td>0.23</td>
</tr>
<tr>
<td>High</td>
<td>0.80</td>
<td>Very low</td>
<td>Adequately powered RCT with little bias</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Empiric evidence of misleading research findings

Study of highly cited studies in high impact journals
- 45 of 49 studies claimed an intervention was effective
- 16% contradicted by subsequent studies
- 16% reported effects stronger than subsequent studies
- 44% were replicated
- 24% had no attempt at replication
- Likelihood of being contradicted or reporting stronger effects
  - Nonrandomized studies: 5 of 6
  - RCTs: 9 of 39 (more likely if smaller samples)

When are positive research findings more likely to be true?

- Predefined, focused hypothesis testing
- Replicates previous findings
- Larger sample sizes studied
- Larger effects reported
- Study designed and executed to minimize bias
  - Need to critically appraise studies to determine risk of bias
Critical appraisal

- All evidence is not equal
- Bias = any process, effect, or error in the design or conduct of a study that systematically favors one outcome over others
  * Studies with greater risk of bias are more likely to be misleading, typically exaggerate treatment effects
- Critical appraisal involves judging features of individual studies related to risk of bias and considering their place on a study design-based evidence hierarchy

Two aspects of quality

- **Internal validity** = the extent to which the results of a study can be reliably attributed to the intervention under evaluation
- **External validity** = the extent to which the results of a study can be applied to other populations and settings

Why patients may appear to improve with a therapy

- Natural history
- Regression to the mean
  * Treatment typically sought when symptoms are worse
- Hawthorne and other non-specific effects
  * Attention, caring, empathy
- Placebo effect
  * Related to perception of treatment
- Specific or “true” effects of treatment

Hawthorne effect

- “An increase in worker productivity produced by the psychological stimulus of being singled out and made to feel important”
  * Simply being studied results in better outcomes
- First reported after a series of studies conducted at the Western Electrical Company’s Hawthorne Works in Chicago in the 1920’s and 1930’s
- Every change in working condition resulted in increased productivity
  * Improving or reducing lighting resulted in similar effects

Uncontrolled studies

- Often retrospective with limited data on baseline characteristics
- Difficult to identify an inception cohort of patients
- Typically not blinded
- Unable to distinguish specific effects of treatment from non-specific effects
- High potential for publication bias

Estimates of effectiveness from uncontrolled studies

- “Mobilization and manipulation studies claim an 80% success rate.”
- “80% of low back pain patients get immediate relief with epidural blocks.”
- “In the YMCA’s exercise program, 80% improve.”
- “With microcurrent therapy…82% were pain free with 10 treatments.”
- “70-80% of those carefully screened for radicular symptoms benefit from surgery.”

*Deyo RA, Spine 1993;18:2153-2162*
Assessing risk of bias in clinical research studies of interventions
Roger Chou MD

Therapies for LBP thought to be effective based on uncontrolled studies

- Bed rest
- Sacroiliac fusion
- Coccygectomy
- Prolotherapy
- Intradiscal steroid injection
- Etc.

Typical hierarchy of evidence

- Well designed randomized controlled trial
- Well designed non-randomized controlled trial
- Well designed prospective cohort study
- Well designed retrospective cohort study
- Well designed case control study
- Before-after studies, case series, case reports, descriptive studies, basic science studies, etc.

Types of bias

<table>
<thead>
<tr>
<th>Type of bias</th>
<th>Factors associated with increased risk of bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection bias</td>
<td>- Inadequate randomization</td>
</tr>
<tr>
<td></td>
<td>- Inadequate allocation concealment</td>
</tr>
<tr>
<td></td>
<td>- Group dissimilarity at baseline</td>
</tr>
<tr>
<td>Performance bias</td>
<td>- Inadequate blinding of participants and study personnel/providers</td>
</tr>
<tr>
<td></td>
<td>- Differential use of co-interventions</td>
</tr>
<tr>
<td></td>
<td>- Differential or suboptimal compliance across groups</td>
</tr>
<tr>
<td>Detection bias</td>
<td>- Inadequate blinding of outcomes assessors</td>
</tr>
<tr>
<td></td>
<td>- Differential timing of outcome assessments</td>
</tr>
<tr>
<td>Attrition bias</td>
<td>- High or differential loss to follow-up</td>
</tr>
<tr>
<td></td>
<td>- Analysis not intention to treat</td>
</tr>
<tr>
<td>Reporting bias</td>
<td>- Selective reporting</td>
</tr>
</tbody>
</table>

Bias in RCTs

Randomization

- Successful randomization largely eliminates the problem of confounding
  - Even unmeasured or unknown confounders will be equally distributed (given adequate randomization methods and enough patients)
- Requires two elements:
  - Generation of a truly unpredictable allocation sequence
  - Successful implementation of the random allocation sequence
- Successful randomization should result in similar baseline characteristics across groups
### Randomization and studies of TENS

Transcutaneous electrical nerve stimulation for post-operative pain

<table>
<thead>
<tr>
<th>Type of study</th>
<th>Number of studies</th>
<th>Number of studies showing a positive treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonrandomized</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>Randomized</td>
<td>17</td>
<td>2</td>
</tr>
</tbody>
</table>


### Randomization and studies of MI

<table>
<thead>
<tr>
<th>Randomization method</th>
<th>Proportion of studies with prognostic maldistribution</th>
<th>Proportion of studies showing a difference in case-fatality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized and concealed allocation</td>
<td>14%</td>
<td>9%</td>
</tr>
<tr>
<td>Randomized, inadequately concealed allocation</td>
<td>27%</td>
<td>24%</td>
</tr>
<tr>
<td>Nonrandomized</td>
<td>58%</td>
<td>58%</td>
</tr>
</tbody>
</table>

Ref: Chalmers T et al. NEJM 1983;309:1358-61

### Randomization: Empiric evidence

- Studies found that inadequate generation of randomization sequence can exaggerate effects by up to 50%, though estimates vary
  - One study found that when limited to studies with adequate allocation concealment, inadequate randomization exaggerated effects by about 25%
  - Seven studies found that inadequate allocation concealment exaggerated treatment effects an average of 18% (95% CI 5-30%)

### Group similarity at baseline

Given a large enough sample, successful randomization should result in equally distributed baseline characteristics

- Statistically significant differences suggest failure of randomization
  - Due to chance, inadequate randomization/allocation concealment methods, or manipulation of treatment allocation
  - Non-statistically significant differences on important baseline characteristics that are clinically relevant may also be important

### Blinding

- Blinding of participants and study personnel/caregivers
  - Double-dummy designs, sham procedures
  - Reduces bias related to knowledge and expectations of treatments received
  - Not always possible
- Blinding of outcomes assessors
  - Reduces bias related to differential assessment of outcomes
  - Almost always possible

### Blinding: Empiric evidence

- Studies show lack of blinding exaggerates estimates of treatment effects an average of 14%
- May be more important in studies that assess subjective outcomes
  - One study found lack of blinding inflated estimates by 30% in trials with subjective outcomes, no effect in trials with objective outcomes
Assessing risk of bias in clinical research studies of interventions
Roger Chou MD

Loss to follow-up

- Loss to follow-up introduces attrition bias when those lost to follow-up differ from those who remain in the study
- Importance depends on amount of loss to follow-up and degree of difference across groups (e.g., low in one group but high in the other)
- Similar bias can occur if subjects are not lost to follow-up but excluded from analysis due to missing data or other reasons

Intention-to-treat analysis

- Patients may not receive allocated treatments, adhere to assigned therapy, or complete the trial
- Intention-to-treat analysis means analyzing patients according to the group they were assigned to, regardless of which treatments they actually receive or complete
- Failure to perform ITT analysis introduces bias when patients who do not adhere to allocated treatments differ systematically from those who do

Selective outcome reporting

Positive and statistically significant findings are more likely to be reported than negative and non-statistically significant findings

Studies should report all pre-specified outcomes
- 71% of outcomes with p<0.05 fully reported
- 50% of outcomes with p>0.05 fully reported
- 30-50% of primary outcomes changed from protocol to publication

Publication bias:

The preferential publication of studies based on the magnitude and direction of the findings

- If missing studies are systematically different from identified studies, then bias is introduced into the systematic review
  - Published studies are more likely to report positive results, resulting in biased estimates
- Methods are available for assessing the presence and impact of publication bias
  - Funnel plot
  - Egger’s test
  - Trim & fill method

Impact of publication bias

Meta-analysis based on published and registered trials

Combination chemotherapy vs. monotherapy in ovarian cancer

Survival ratio

(95% confidence interval)

Published trials (p=0.004)
Registered trials (p=0.17)

Assumptions of methods for assessing for publication bias

- Large studies are likely to be published regardless of statistical significance
- Small studies are at greatest risk for being lost: studies reported largest effects are more likely to be published
- Effect of these assumptions:
  - Expect bias to increase as sample size goes down
- Various methods for assessing publication bias are based on looking for small sample effects
  - Other causes of small sample effects include differences in methodological quality, chance
Assessing risk of bias in clinical research studies of interventions

Roger Chou MD

**Funnel Plot**
- Scatter plot of sample size or other measure of variance on y-axis, effect size on x-axis
- Smaller studies show greater variability in effects, plot should appear like inverted funnel
  - If smaller studies with larger effects are preferentially published, the base of the funnel will appear skewed

**Funding source**
Can be an indirect source of bias
- Many studies show that studies funded by industry tend to report more favorable results for the funder’s products, potentially via:
  - Selective publication of favorable studies
  - Selective reporting of favorable outcomes
  - “Spinning” of results to make them seem more favorable
  - Designing studies to favor one intervention over another
  - Comparing against a less effective intervention
  - Selecting patients more likely to be responsive
  - Excluding patients who may experience side effects

**Observational studies**
- Many types of observational studies
- Cohort and case-control studies often used to evaluate effects of medical interventions
  - Have separate comparison groups
  - Highly susceptible to confounding
  - Require additional safeguards not needed in well-designed RCTs
  - Confounding by indication is prominent in studies of treatment efficacy
  - Residual confounding likely to be present even if techniques (matching, restriction, adjustment) used to minimize it
  - Tend to have less control over interventions, co-interventions, adherence, assessment of outcomes, and follow-up

**Limitations of matching**

<table>
<thead>
<tr>
<th></th>
<th>Apples</th>
<th>Oranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape</td>
<td>Round</td>
<td>Round</td>
</tr>
<tr>
<td>Source</td>
<td>Tree</td>
<td>Tree</td>
</tr>
<tr>
<td>Edible?</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Size</td>
<td>Handheld</td>
<td>Handheld</td>
</tr>
<tr>
<td>Weight</td>
<td>½ lb</td>
<td>½ lb</td>
</tr>
</tbody>
</table>

**When should clinicians adopt research findings into practice?**
- Caution with initial, unreplicated findings
  - Protasis effects—tendency for initial publication of positive findings, followed by extreme contradictory results
  - Publication bias
- Caution with findings from small studies
- Caution when reported effect sizes are small
- Critically appraise studies to assess reliability of findings
- Consider looking to a well-done systematic review or guideline that summarizes all of the available evidence
Assessing risk of bias in clinical research studies of interventions
Roger Chou MD

Evaluating a body of evidence: GRADE method

- Risk of bias
  - Study design
  - Aggregate quality of the studies
- Consistency of results
- Effect sizes are in the same direction
  - The range of effect sizes is narrow
- Precision of estimates
  - Degree of certainty regarding an effect estimate with respect to a given outcome
- Directness of evidence
  - Health outcomes rather than intermediate outcomes
  - Direct comparisons rather than indirect comparisons

Be prepared to re-think practices!

Negative findings are often more likely to be true than positive findings, yet:

- Contradicted positive claims often persist for many years despite convincing negative evidence from RCT’s
- 50% of articles on vitamin E still reported favorable association with CV risk 8 years after publication of a major contradictory RCT
- Beta-carotene and cancer, estrogen and Alzheimer’s disease
- Failure or unwillingness to abandon established medical practices in the face of convincing negative evidence
  - Intensive insulin therapy for ICU patients

Discrepancy between early, small trials and later, larger trials

<table>
<thead>
<tr>
<th>Study</th>
<th>Odds ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head to Head</td>
<td></td>
</tr>
<tr>
<td>Intensive insulin therapy for ICU patients:</td>
<td></td>
</tr>
</tbody>
</table>

CVA or death after CEA with patching vs. primary closure

Consistent results across higher-quality RCTs is critical

- Reproducibility of results is a core principle of the scientific method
  - Findings that are reproduced increase the pre-test probability of subsequent studies
- Magnesium for myocardial infarction
  - 7 trials with 1300 patients showed benefit
  - Subsequent large trial (ISIS-4) showed no benefit
  - "Several medium-sized trials of high quality seem necessary to render results trustworthy" (Egger and Davey Smith BMJ 1995;310:752-4)

Example: Glucose control in SICU patients

- Trial by van der Bergh et al showed decreased mortality with tight glucose control in SICU patients (NEJM 2001)
  - N=1548
  - Adopted in many guidelines and settings
- Five subsequent trials showed no benefit on mortality (1 trial showed increased mortality) and increased risk of hypoglycemia

Finding What Works In Health Care

“...The most reliable way to identify benefits and harms associated with various treatment options is a systematic review of comparative effectiveness research.”

The Institute of Medicine, 2011
Why do we need reviews and guidelines?

- Proliferation of medical literature
- Individual studies can be misleading, or provide conflicting findings
  - Variable design and quality of primary studies
- Complex links between interventions and outcomes may not be captured in individual studies
- Individual studies may lack sufficient power to detect small but clinically important effects
- Helps build upon previous research
  - What questions are unanswered?
  - What research would answer these questions?

Proliferation of medical literature

- More RCTs were run in the year 2000, than in the entire decade 1965-1975
- Doubling time of biomedical science:
  - ~19 years in 1991
  - ~20 months in 2001
- MEDLINE added 683,136 new records in 2008:
  - 13,137/week
  - 1,866/day

Traditional review articles

- Do not use systematic literature searches
- Do not describe how studies were chosen for inclusion
- Do not assess or use reproducible methods to assess the validity of included studies
- Evidence is often incomplete or influenced by biases or opinions of authors
- Frequently recommend treatments long after they have been shown to be useless, or even harmful

The Systematic Review

“In its ideal form, is a review that includes an explicit and detailed description of how it was conducted so that any interested reader would be able to replicate it”

(Systematic reviews)

- Are systematic to remove bias in finding and reviewing the literature
- Use explicit methods for:
  - Identifying all relevant studies
  - Including or excluding studies
  - Rating the validity of studies
  - Synthesizing the evidence, either qualitatively or quantitatively (‘meta-analysis’)
**Assessing risk of bias in clinical research studies of interventions**

Roger Chou MD

### Uses of systematic reviews

- Summarize the existing data
- Detect small but clinically meaningful treatment effects
- Evaluate consistency of results between studies
- Develop and test hypotheses to explain differential outcomes between studies
- Identify research gaps and areas of uncertainty
- For decision-makers, provide an objective framework for collecting and reviewing evidence to guide policy decisions

### Clinical practice guidelines

- Another form of pre-processed evidence that can help guide clinical practice
- Should be based on a systematic review of the evidence
- Go beyond systematic reviews by providing recommendations to guide clinical practice, not just describe the evidence
- Should have an explicit link between the evidence and the recommendations
- Conflicts of interest are critical as they can affect the judgments used to develop recommendations
- Institute of Medicine standards for guidelines issued in 2011

### Conclusions

- Positive findings from research studies are often misleading
- Factors associated with a higher likelihood of falsely positive findings include small sample sizes, low pretest probability, and risk of bias
- Critical appraisal is vital!
- Reproducibility of findings is a core scientific principle!
- Don’t ignore negative results!
- Consider the whole body of evidence before making judgments
- Systematic reviews and clinical practice guidelines can be useful for guiding clinical practice, but must also adhere to methodological standards to be trustworthy