Getting Started With Systematic Reviews

W. Scott Richardson, M.D.
AU/UGA Partnership, Athens
Three Owl Learning Institute
Conflicts of Interest

• I have no financial ties with industry that pose a conflict of interest regarding the content of this presentation

• I will not be discussing “off label” uses of any medications or devices

• Image copyrights are retained by their original creators, publishers, etc.
Session Aims

Structure and function of systematic reviews of treatment trials

Appraise SR methods

Understand SR results
EBM: Why Bother

• We can’t make well-informed decisions without information
• Not all information is created equal
• Misinformation can be worse than no information
• Better information → better informed decisions → better outcomes
Individual randomized trials of treatment ...

- Each trial is one experiment, one new chance to get closer to the ‘truth’
- One trial ~ one race
- Often, more than one trial is done
- Will all trial results agree (even by chance)?
As trials accumulate …

- Seldom is one trial definitive ("One ring to rule them all …")

- In science, as experiments accrue, knowledge is built cumulatively

- Is there a scientific way to combine results of individual trials?

- Yes! Systematic reviews (we’ll abbreviate “SRs”)
‘Narrative’ vs. ‘Systematic’

- Address disorder as a whole – overview
- Or, tell a ‘story’
- Variety of questions
- No methods section
- No formal pooling
- Thus, may be cumulative but not comprehensive

- Address focused question (e.g. effect of therapy, accuracy of diagnostic test)
- Methods section
- Formal pooling, when appropriate
- Thus, cumulative and comprehensive
SR Methods

• Formulate questions
• Define eligibility criteria for study inclusion
• Develop *a priori* hypotheses to explain heterogeneity
• Conduct search
• Screen titles, abstracts for inclusion, exclusion
• Review full text

• Assess the risk of bias
• Abstract data
• When meta-analysis is performed:
  – Summary estimates, confidence intervals
  – Explain heterogeneity
  – Rate confidence in estimates of effect
• Report results
• Update review as needed
‘PRISMA’

• ‘Preferred Reporting Items for Systematic reviews and Meta-Analyses’
• Incorporates evolutionary advances
• Specifies 27 item checklist for reporting, e.g. standardizes figures, etc.
• Since 2009, has replaced ‘QUOROM’, has been adopted by many journals
• Ann Intern Med 2009; 151: 264 – 269
Finding SRs

• Cochrane Library
  – CDSR – Cochrane Database of Systematic Reviews
  – DARE – Database of Abstracts of Reviews of Effects

• PubMed
  – Publication types
  – Clinical queries

• Work with your team to find SRs
Critical Appraisal of SRs

Credibility:
- Sensible question?
- Exhaustive search?
- Selection, assessments reproducible?
- Present results ready for application?
- Address confidence in estimates of effect?

Confidence in Estimates:
- Risk of bias?
- Consistent across studies?
- Effect: RR, OR, WMD
- Precision: 95% CI
- Apply to my patient?
- Reporting bias?
- Reasons to increase confidence rating?
‘Risk of bias’

- Moves away from dichotomous “yes/no” to explicit rating of risk of bias
- At both study-level and outcome-level
- *BMJ* 2011; 343: d5928 doi
Risk of bias graphs

Figure 1. Risk of bias graph: review authors' judgements about each risk of bias item presented as percentages across all included trials.
Reporting Biases

• Selective reporting of studies
  – Delayed (or never)
  – Location, language
• Selective reporting of outcomes, times
• Selective reporting of analyses
• UG 3/e Box 23-2

• Empirical evidence
• Distort the ‘body of evidence’ in the literature
• Can lead to wrong conclusions about the benefits and harms
Pyramid vs GRADE

A

RCT
Cohort study
Case control study
Case series
Case reports
Animal research
In-vitro research
Expert experience

B

<table>
<thead>
<tr>
<th>Quality of evidence</th>
<th>Study design</th>
<th>Lower quality if*</th>
<th>Higher quality if†</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Randomised trial</td>
<td>Study limitations - 1 serious - 2 very serious</td>
<td>Large effect + 1 large + 2 very large</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inconsistency - 1 serious - 2 very serious</td>
<td>Dose response + 1 evidence of a gradient</td>
</tr>
<tr>
<td>Moderate</td>
<td></td>
<td>Indirectness - 1 serious - 2 very serious</td>
<td>All plausible confounders + Would reduce a demonstrated effect or + Would suggest a spurious effect when results show no effect</td>
</tr>
<tr>
<td>Low</td>
<td>Observational study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very low</td>
<td></td>
<td>Imprecision - 1 serious - 2 very serious</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Publication bias - 1 likely - 2 very likely</td>
<td></td>
</tr>
</tbody>
</table>
### Forest Plot – Incidence of recurrent VTE

<table>
<thead>
<tr>
<th>Study or subgroup</th>
<th>LMWH</th>
<th>VKA</th>
<th>Odds Ratio</th>
<th>Weight</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n/N</td>
<td>n/N</td>
<td>Peto, Fixed</td>
<td>95% CI</td>
<td>Peto, Fixed</td>
</tr>
<tr>
<td>Das 1996</td>
<td>5/50</td>
<td>2/55</td>
<td>2.75 [0.60, 12.69]</td>
<td>5.5 %</td>
<td>2.75 [0.60, 12.69]</td>
</tr>
<tr>
<td>Daskalopoulos 2005</td>
<td>2/50</td>
<td>3/52</td>
<td>0.69 [0.11, 4.11]</td>
<td>4.0 %</td>
<td>0.69 [0.11, 4.11]</td>
</tr>
<tr>
<td>Gonzalez 1999</td>
<td>8/93</td>
<td>19/92</td>
<td>0.38 [0.17, 0.86]</td>
<td>19.4 %</td>
<td>0.38 [0.17, 0.86]</td>
</tr>
<tr>
<td>Hamann 1998</td>
<td>3/100</td>
<td>2/100</td>
<td>1.50 [0.26, 8.84]</td>
<td>4.1 %</td>
<td>1.50 [0.26, 8.84]</td>
</tr>
<tr>
<td>Hull 2007</td>
<td>18/369</td>
<td>21/368</td>
<td>0.85 [0.45, 1.62]</td>
<td>30.9 %</td>
<td>0.85 [0.45, 1.62]</td>
</tr>
<tr>
<td>Kakkas 2003</td>
<td>3/103</td>
<td>5/221</td>
<td>1.31 [0.29, 5.89]</td>
<td>5.7 %</td>
<td>1.31 [0.29, 5.89]</td>
</tr>
<tr>
<td>Lopaciuk 1999</td>
<td>3/101</td>
<td>7/101</td>
<td>0.43 [0.12, 1.54]</td>
<td>8.0 %</td>
<td>0.43 [0.12, 1.54]</td>
</tr>
<tr>
<td>Lopez 2001</td>
<td>0/81</td>
<td>2/77</td>
<td>0.13 [0.01, 1.22]</td>
<td>2.5 %</td>
<td>0.13 [0.01, 1.22]</td>
</tr>
<tr>
<td>Pini 1994</td>
<td>6/93</td>
<td>4/94</td>
<td>1.54 [0.43, 5.49]</td>
<td>8.0 %</td>
<td>1.54 [0.43, 5.49]</td>
</tr>
<tr>
<td>Romeroa 2009</td>
<td>5/119</td>
<td>7/122</td>
<td>0.72 [0.23, 2.31]</td>
<td>9.6 %</td>
<td>0.72 [0.23, 2.31]</td>
</tr>
<tr>
<td>Veiga 2000</td>
<td>2/50</td>
<td>1/50</td>
<td>1.97 [0.20, 19.43]</td>
<td>2.5 %</td>
<td>1.97 [0.20, 19.43]</td>
</tr>
<tr>
<td><strong>Total (95% CI)</strong></td>
<td><strong>1209</strong></td>
<td><strong>1332</strong></td>
<td><strong>100.0 %</strong></td>
<td><strong>0.77 [0.54, 1.10]</strong></td>
<td></td>
</tr>
</tbody>
</table>

Total events: 55 (LMWH), 74 (VKA)
Heterogeneity: Chi² = 11.68, df = 10 (P = 0.31); I² = 14%
Test for overall effect Z = 1.44 (P = 0.15)
Test for subgroup differences: Not applicable
### Forest plot – b

#### Deaths/Total

<table>
<thead>
<tr>
<th></th>
<th>Albumin</th>
<th>Saline</th>
</tr>
</thead>
<tbody>
<tr>
<td>All patients</td>
<td>644/3012</td>
<td>655/3028</td>
</tr>
<tr>
<td>Baseline serum albumin concentration ≤25 g/l</td>
<td>291/1228</td>
<td>321/1223</td>
</tr>
<tr>
<td>Baseline serum albumin concentration &gt;25 g/l</td>
<td>353/1784</td>
<td>334/1805</td>
</tr>
<tr>
<td>Heterogeneity P=0.08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Odds ratio (95% CI)

- **All patients**: 0.99 (0.87 to 1.11)
- **Baseline serum albumin concentration ≤25 g/l**: 0.87 (0.73 to 1.05)
- **Baseline serum albumin concentration >25 g/l**: 1.09 (0.92 to 1.28)

---

**Fig 5** | Forest plot from study comparing resuscitation with albumin or saline in intensive care showing unadjusted odds ratio of death stratified by baseline albumin concentration.

---
Are you happy pooling?

Relative Risk (95% CI)

- 0.73 (0.49, 1.07)
- 0.74 (0.59, 0.94)
- 0.76 (0.51, 1.12)
- 0.71 (0.56, 0.90)
- 0.73 (0.61, 0.88)
Are you happy pooling?

Relative Risk (95% CI)

- 0.44 (0.30, 0.65)
- 0.45 (0.36, 0.60)
- 1.25 (0.84, 1.84)
- 1.17 (0.92, 1.49)
- 0.73 (0.61, 0.88)
What criteria were you using?

• similarity of point estimates
  – less similar, less happy

• overlap of confidence intervals
  – less overlap, less happy
Heterogeneity

- Humans vary, e.g. in risk of poor outcomes from disease, in response to therapy, and in vulnerability to adverse effects
- Heterogeneity represents this variation in results
- Affects certainty about estimates of effect

- Identified by:
  - Visual inspection
  - Chi^2: “yes” or “no”
  - I^2: 0 to 100%

- Explored by:
  - Patients
  - Interventions
  - Comparisons
  - Outcomes
  - Methods, Systems, +
Homogenous

Ho: RR1 = RR2 = RR3 = RR4

test for heterogeneity
what is the p-value?

Relative Risk (95% CI)
0.73 (0.49, 1.07)
0.74 (0.59, 0.94)
0.76 (0.51, 1.12)
0.71 (0.56, 0.90)
0.73 (0.61, 0.88)

p=0.99 for heterogeneity
Heterogeneous

Relative Risk (95% CI)

0.44 (0.30, 0.65)
0.45 (0.36, 0.60)
1.25 (0.84, 1.84)
1.17 (0.92, 1.49)
0.73 (0.61, 0.88)

p-value for heterogeneity < 0.001

test for heterogeneity

what is the p-value?
I² Interpretation

100%
Why are we pooling?

Very concerned

Getting concerned

Only a little concerned

0%
No worries
Homogenous

What is the $I^2$?

$p=0.99$ for heterogeneity

$I^2=0\%$
Heterogeneous

What is the $I^2$?

$p$-value for heterogeneity < 0.001

$I^2$=89%
Why Not Use Subgroups?

• Subgroups may be informative for clinical decisions (in present) and raise hypothesis for further research (in the future)

• Subgroups may also mislead, due to several possible explanations for differences found

Possible explanations of difference in subgroups:
• Hypothesized difference
• Chance
• Other patient difference
• Different co-interventions
• Different outcome measures
• Different risk of bias
Multiple looks; imbalance

- If no difference exists, multiple comparisons risks finding ‘false positive’ results.
- “The more you look, the more you find.”
- Using subgroups undoes the prognostic balance from random allocation.
Credibility of subgroup analyses

Ten criteria in 3 main areas
  – Study design
  – Data analysis
  – Study context

Greater confidence if most or all are met

Lower confidence if few or none are met

Work through with teams during appraisal

<table>
<thead>
<tr>
<th>Criteria to assess the credibility of subgroup analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Design</strong></td>
</tr>
<tr>
<td>• Is the subgroup variable a characteristic measured at baseline or after randomisation? *</td>
</tr>
<tr>
<td>• Is the effect suggested by comparisons within rather than between studies?</td>
</tr>
<tr>
<td>• Was the hypothesis specified a priori?</td>
</tr>
<tr>
<td>• Was the direction of the subgroup effect specified a priori *</td>
</tr>
<tr>
<td>• Was the subgroup effect one of a small number of hypothesised effects tested?</td>
</tr>
<tr>
<td><strong>Analysis</strong></td>
</tr>
<tr>
<td>• Does the interaction test suggest a low likelihood that chance explains the apparent subgroup effect?</td>
</tr>
<tr>
<td>• Is the significant subgroup effect independent? *</td>
</tr>
<tr>
<td><strong>Context</strong></td>
</tr>
<tr>
<td>• Is the size of the subgroup effect large?</td>
</tr>
<tr>
<td>• Is the interaction consistent across studies?</td>
</tr>
<tr>
<td>• Is the interaction consistent across closely related outcomes within the study? *</td>
</tr>
<tr>
<td>• Is there indirect evidence that supports the hypothesised interaction (biological rationale)?</td>
</tr>
</tbody>
</table>

*New criteria.
SR’s of Other Study Types

- Diagnostic test accuracy studies
- Cohort studies of prognosis
- Disease probability for differential diagnosis*
- Other observational studies
How quickly do systematic reviews go out of date?

- Survival analysis
- 100 systematic reviews, 1995 – 2005
- Searched for ‘update signals’ (i.e. new trial evidence)
- **Ann Intern Med 2007**

The immediate decrease in survival at time zero reflects the 7 systematic reviews for which signals for updating had already occurred at the time of publication. The low number of reviews at risk after 10 years reflects the fact that the sample spanned 1995 to 2005 and censoring occurred on 1 September 2006. Thus, only reviews published before September 1996 and having no signals for updating could have more than 10 years of observation.
Taking SRs home …

• When well-made and current, SRs synthesize the body of research evidence that can guide important decisions

• SRs have limits, yet we should start with them: ‘how well does this work?’

• We can (and must!) appraise SRs for risk of bias, estimates of effect, and confidence in these estimates
Questions?