Reliability of Three Assessment Tools Used to Evaluate Randomized Controlled Trials for Treatment of Neck Pain

N Graham MSc, PT  
T Haines MSc, MD  
C Goldsmith PhD

U Shahzad BHSc (Hon)  
E Talovikova BHSc (Hon)  
A Gross MSc, PT  
S Burnie MSc, DC

for Cervical Overview Group (COG)
Cervical Overview Group
1992 to 2010
Overview

- Brief overview of COG and Validity team
- Objective
- Background Summary
- Methods
- Results
- Conclusions
COG
1992 to 2010

Medicine
- Injections
- Medication

Physical Medicine Methods
- LLLT
- Traction
- Electrotherapy
- Heat & Cold
- Ultrasound
- Exercise
- Acupuncture
- Orthosis

Manual Therapy
- Manipulation
- Mobilization
- Massage

Patient Education
Standardized method for systematic review

- Literature Search
  - Identification criteria
  - Selection criteria
- Validity Assessment
- Data Abstraction
  - Analysis: Effect Measure
  - Test of Homogeneity
- Sensitivity Analysis
- Synthesis
- Conclusions
- Recommendations
Objective

- To assess the inter-rater reliability of three tools used by COG for assessment of internal validity of RCTs

Design: Pragmatic, cross-sectional study
June 2003 – June 2009

Tools: Jadad, van Tulder & Risk of Bias
Background

- Numerous tools exist
  - Jadad
  - PEDro
  - Delphi List
  - van Tulder (CBRG method guidelines)
  - Cochrane Risk of Bias (CBRG method guidelines and Cochrane Handbook Chapter 8)

- Scales and checklists

- Criteria-based domains
Jadad

- Developed through standardized item reduction process in 1996
- 3 item scale
  - Randomization, blinding, drop-outs
- Total score 5
- Score of 3 or more indicative of high quality
- Validated for evaluation of pain in drug trials

Jadad et al 1996
## JADAD SCORING CRITERIA

<table>
<thead>
<tr>
<th>JADAD SCORING CRITERIA</th>
<th>Potential Score</th>
<th>Score Awarded</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. Was the study described as randomized?</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>1b. Was the method of randomization described and appropriate to conceal allocation?</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>1c. If described and inappropriate, describe:</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>2a. Was the study described as double blinded?</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>2b. Was the method of double-blinding described and appropriate to maintain a double-blinding?</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td>2c. Was the method of blinding inappropriate?</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>3. Was there a description of withdrawals and drop outs?</td>
<td>+1</td>
<td></td>
</tr>
<tr>
<td><strong>FINAL SCORE (0 – 5)</strong></td>
<td>5</td>
<td>__ / 5</td>
</tr>
</tbody>
</table>
van Tulder

- Checklist developed in 1997 by the CBRG editorial board as recommended method guidelines for systematic reviews in area of spinal disorders
- Updated in 2003 revised to 11 item criteria list for methodological quality assessment
  - 6/11 considered high quality
- Updated in 2009 revised to 12 item internal validity checklist for source of risk of bias
  - recent evidence supporting the use of sum scores

<table>
<thead>
<tr>
<th>van Tulder 11 item checklist (2003)</th>
<th>Yes</th>
<th>No / DK</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Was the method of randomization adequate?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. Was the treatment allocation concealed?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Were the groups similar at baseline regarding the most important prognostic indicators?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. Was the patient blinded to the intervention?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. Was the care provider blinded to the intervention?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F. Was the outcome assessor blinded to the intervention?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G. Were co-interventions avoided or similar?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H. Was the compliance acceptable in all groups?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I. Was the withdrawal/drop-out rate described &amp; acceptable?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J. Was the timing of outcome assessment in all grps similar?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K. Did the analysis include an intention-to-treat analysis?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TOTAL ITEMS SCORED YES (0 - 11)** __ / 11
Risk of Bias

- Cochrane Collaboration recommends domain based criteria
  - Sequence generation
  - Allocation concealment
  - Blinding of participants, personnel and outcome assessors
  - Incomplete outcome data
  - Selective outcome reporting
  - Other sources of bias

- Not a scale or checklist with sum score
- Qualitative assessment requiring subjective judgment

Cochrane Handbook 2008, Chapter 8
## Risk of Bias

| A | **Sequence generation**  
1. Was the method of randomization adequate? | Yes / No / Unsure |
|---|---|---|
| B | **Allocation concealment**  
2. Was the treatment allocation concealed? | Yes / No / Unsure |
| C | **Blinding of participants, personnel and outcome assessors**  
Was knowledge of the allocated interventions adequately prevented during the study?  
3. Was the patient blinded to the intervention?  
4. Was the care provider blinded to the intervention?  
5. Was the outcome assessor blinded to the intervention? | Yes / No / Unsure |
| D | **Incomplete outcome data**  
*Were incomplete outcome data adequately addressed?*  
6. Was the drop-out rate described and acceptable?  
7. Were all randomized participants analyzed in the group to which they were allocated? | Yes / No / Unsure |
| E | **Selective outcome reporting**  
8. Are reports of the study free of suggestion of selective outcome reporting? | Yes / No / Unsure |
| F | **Other sources of potential bias**  
9. Were the groups similar at baseline regarding the most important prognostic indicators?  
10. Were co-interventions avoided or similar?  
11. Was the compliance acceptable in all groups?  
12. Was the timing of the outcome assessment similar in all groups? | Yes / No / Unsure |

Furlan et al 2009, CBRG
Methods

- Four members of COG, multi-professional and methodological backgrounds

- Evaluation of internal validity of 54 RCTs using Jadad and van Tulder, 18 RCTs using RoB from June 2003 to June 2009

- Kappa statistic with standard agreement categorization
Methods

- Each rater (Statistician, MD, PT, DC) independently assessed RCT using all 3 tools
- Meeting to discuss, consensus reached
- Individual and consensus recorded
- Kappa statistic calculated for each combination of raters
Methods

Scale to interpret Kappa score

- $\leq 0 = \text{poor}$
- $0.01$ to $0.2 = \text{slight}$
- $0.21$ to $0.4 = \text{fair}$
- $0.41$ to $0.6 = \text{moderate}$
- $0.61$ to $0.8 = \text{substantial}$
- $0.81$ to $1 = \text{almost perfect}$

Landis and Koch 1977
Results

- Jadad 4/7 moderate to substantial agreement
- van Tulder 8/11 moderate agreement
- Risk of Bias 3/12 moderate to substantial
Results

Kappa mean (min to max)

- Consistent substantial agreement across all tools for the domain **allocation concealment**
  - Jadad 0.69 (0.60 to 0.77)
  - van Tulder 0.77 (0.73 to 0.81)
  - Risk of Bias 0.76 (0.65 to 0.88)

- Consistent substantial agreement across two tools for the domain **randomization**
  - van Tulder 0.53 (0.37 to 0.66)
  - Risk of Bias 0.66 (0.45 to 0.88)

- Other domains demonstrated fair to moderate agreement or were not a fair test of agreement
# Results Jadad (n=54)

Kappa estimate (95% CI: lower bound, higher bound)

<table>
<thead>
<tr>
<th>R*</th>
<th>1b Appropriate concealment</th>
<th>1c Inappropriate concealment</th>
<th>2a Double blinding</th>
<th>2b Appropriate blinding</th>
<th>3a Dropouts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>0.77, 0.94</td>
<td>0.48, 1.00</td>
<td>0.80, 1.00</td>
<td>0.65, 0.93</td>
<td>0.21, -0.05, 0.47</td>
</tr>
<tr>
<td>1, 3</td>
<td>0.60, 0.82</td>
<td>0.46, 0.90</td>
<td>0.73, 0.98</td>
<td>0.34, 0.700</td>
<td>0.16, -0.08, 0.42</td>
</tr>
<tr>
<td>2, 3</td>
<td>0.69, 0.88</td>
<td>0.31, 0.78</td>
<td>0.80, 1.00</td>
<td>0.61, 0.91</td>
<td>0.09, -0.18, 0.37</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>0.42</td>
<td>0.78</td>
<td>0.53</td>
<td>0.15</td>
</tr>
</tbody>
</table>

R* = raters

**1a (randomization) and 2c (inappropriate blinding) were not computed**
## Results van Tulder (n=54)
### Item A to F

<table>
<thead>
<tr>
<th>R*</th>
<th>A - Randomization adequate</th>
<th>B - Allocation concealment</th>
<th>C - Groups at baseline</th>
<th>D - Patient blinding</th>
<th>E - Care provider blinding</th>
<th>F - Assessor blinding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>0.53</td>
<td>0.81</td>
<td>0.42</td>
<td>0.65</td>
<td>0.26</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>0.30,0.76</td>
<td>0.65,0.96</td>
<td>0.17,0.66</td>
<td>0.42,0.88</td>
<td>-0.15,0.67</td>
<td>0.42,0.85</td>
</tr>
<tr>
<td>1, 3</td>
<td>0.37</td>
<td>0.77</td>
<td>0.40</td>
<td>0.72</td>
<td>0.64</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>0.14,0.60</td>
<td>0.60,0.94</td>
<td>0.16,0.65</td>
<td>0.50,0.94</td>
<td>0.27,1.00</td>
<td>0.25,0.70</td>
</tr>
<tr>
<td>2, 3</td>
<td>0.66</td>
<td>0.73</td>
<td>0.50</td>
<td>0.54</td>
<td>0.48</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>0.49,0.86</td>
<td>0.55,0.91</td>
<td>0.26,0.74</td>
<td>0.27,0.82</td>
<td>-0.11,1.00</td>
<td>0.12,0.63</td>
</tr>
<tr>
<td>Mean</td>
<td>0.53</td>
<td>0.77</td>
<td>0.44</td>
<td>0.64</td>
<td>0.46</td>
<td>0.50</td>
</tr>
</tbody>
</table>

R* = raters

Kappa estimate (95% CI: lower bound, higher bound)
## Results van Tulder (n=54)
### Item G to K

<table>
<thead>
<tr>
<th>R*</th>
<th>G Co-interventions</th>
<th>H Acceptable compliance</th>
<th>I Dropouts</th>
<th>J Timing of Assessment</th>
<th>K Intention-to-treat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>0.23 -0.02,0.48</td>
<td>0.33 0.09,0.58</td>
<td>0.29 0.01,0.58</td>
<td>-0.02 -0.07,0.01</td>
<td>0.44 0.20,0.68</td>
</tr>
<tr>
<td>1, 3</td>
<td>0.29 0.03,0.54</td>
<td>0.52 0.29,0.74</td>
<td>0.39 0.11,0.67</td>
<td>-0.02 -0.06,0.01</td>
<td>0.63 0.42,0.83</td>
</tr>
<tr>
<td>2, 3</td>
<td>0.58 0.37,0.79</td>
<td>0.51 0.28,0.74</td>
<td>0.08 -0.20,0.37</td>
<td>0.79 0.39,1.00</td>
<td>0.58 0.37,0.80</td>
</tr>
<tr>
<td>Mean</td>
<td>0.36</td>
<td>0.45</td>
<td>0.26</td>
<td>0.24</td>
<td>0.55</td>
</tr>
</tbody>
</table>

R* = raters
# Results RoB (n=18) Item 1 to 7

Kappa estimate (95% CI: lower bound, higher bound)

<table>
<thead>
<tr>
<th>R*</th>
<th>1 Randomization</th>
<th>2 Allocation concealment</th>
<th>3 Patient blinding</th>
<th>5 Assessor blinding</th>
<th>6 Dropouts</th>
<th>7 Intention-to-treat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 4</td>
<td>0.88 0.67,1.00</td>
<td>0.65 0.30,1.00</td>
<td>0.15 -0.30,0.60</td>
<td>0.26 -0.16,0.69</td>
<td>0.44</td>
<td>-0.05 -0.48,0.37</td>
</tr>
<tr>
<td>1, 2</td>
<td>0.45 0.07,0.84</td>
<td>0.77 0.48,1.00</td>
<td>0.36 -0.11,0.85</td>
<td>0.36 -0.11,0.85</td>
<td>0.55</td>
<td>0.41 0.01,0.82</td>
</tr>
<tr>
<td>1, 3</td>
<td>0.76 0.46,1.00</td>
<td>0.88 0.67,1.00</td>
<td>N/C</td>
<td>N/C</td>
<td>0.33</td>
<td>0.44 -0.01,0.90</td>
</tr>
<tr>
<td>2, 4</td>
<td>0.56 0.19,0.93</td>
<td>0.88 0.67,1.00</td>
<td>0.76 0.34,1.00</td>
<td>0.45 -0.14,1.00</td>
<td>0.65</td>
<td>0.10 -0.36,0.56</td>
</tr>
<tr>
<td>3, 4</td>
<td>0.88 0.67,1.00</td>
<td>0.76 0.46,1.00</td>
<td>N/C</td>
<td>N/C</td>
<td>0.20</td>
<td>-0.05 -0.48,0.37</td>
</tr>
<tr>
<td>2, 3</td>
<td>0.45 0.07,0.84</td>
<td>0.65 0.30,1.00</td>
<td>N/C</td>
<td>N/C</td>
<td>0.06</td>
<td>0.64 0.31,0.98</td>
</tr>
<tr>
<td>Mean</td>
<td>0.66 0.76</td>
<td>0.76 0.42</td>
<td>0.42</td>
<td>0.35</td>
<td>0.37</td>
<td>0.24</td>
</tr>
</tbody>
</table>

R* = Reviewers; **4 (care provider blinding) was removed; N/C = not computed
### Results RoB (n=18) Item 8 to 12

**Kappa (95% CI: lower bound, higher bound)**

<table>
<thead>
<tr>
<th>R*</th>
<th>8 Selective reporting</th>
<th>9 Groups similar at baseline</th>
<th>10 Co-interventions</th>
<th>11 Compliance</th>
<th>12 Timing of outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 4</td>
<td>-0.15 -0.30, 0.00</td>
<td>0.40 -0.02, 0.83</td>
<td>0.49 0.03, 0.94</td>
<td>0.41 0.01, 0.82</td>
<td>N/C</td>
</tr>
<tr>
<td>1, 2</td>
<td>0.30 -0.28, 0.89</td>
<td>0.10 -0.31, 0.52</td>
<td>0.40 -0.01, 0.81</td>
<td>0.41 0.01, 0.82</td>
<td>-0.05 -0.14, 0.02</td>
</tr>
<tr>
<td>1, 3</td>
<td>0.30 -0.28, 0.89</td>
<td>0.65 0.32, 0.98</td>
<td>0.60 0.13, 1.00</td>
<td>0.16 -0.31, 0.65</td>
<td>-0.05 -0.14, 0.02</td>
</tr>
<tr>
<td>2, 4</td>
<td>0.43 -0.21, 1.00</td>
<td>0.33 -0.09, 0.75</td>
<td>0.87 0.62, 1.00</td>
<td>0.32 -0.11, 0.76</td>
<td>N/C</td>
</tr>
<tr>
<td>3, 4</td>
<td>0.43 -0.21, 1.00</td>
<td>0.55 0.18, 0.93</td>
<td>0.55 0.11, 0.99</td>
<td>0.41 0.01, 0.82</td>
<td>N/C</td>
</tr>
<tr>
<td>2, 3</td>
<td>1.00 0.29 -0.15, 0.74</td>
<td>0.45 0.01, 0.89</td>
<td>0.41 0.01, 0.82</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.38 0.38 0.56</td>
<td>0.35 0.30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R* = Reviewers; **4 (care provider blinding) was removed; N/C= not computed
Conclusions

- Blinding is inherent limitation in rehab trials
- Incomplete reporting, use of “unsure” leads to greater variation of kappa scores
- Decision rules change and new members need to be calibrated
Conclusions

- Diverse group, generalizability, fairer test of scoring reliability
- Consistent inter-rater agreement for allocation concealment and randomization
- Some items require more judgment
Thank you

Acknowledgement
McMaster University
Carey Chittley, Lisa Ditchburn, Danna Epstein, Denise Ward, Guillaume Bourrillon