Holger Schünemann
Professor and Chair, Dept. of Clinical Epidemiology & Biostatistics
Professor of Medicine
Michael Gent Chair in Healthcare Research
McMaster University, Hamilton, Canada

Berlin, 31 Januar, 2013
Methodische Besonderheiten, Rahmenbedingungen und Herausforderungen bei Studien mit nicht-medikamentösen und nicht-technischen Behandlungsverfahren
Disclosure

• No direct/personal for-profit payments to me or my research group
• Co-chair of GRADE working group
• Cochrane Collaboration
  – Co-convenor of the Applicability and Recommendations Methods Group
  – Various other functions
• IQWiG Scientific Board
• Methodological idiosyncrasies
  – Non-pharmaceutical, non-technical interventions (NPNTI) = context of the simple to complex intervention continuum
  – Studies often require cluster or expertise-based randomized designs (design/sample size & directness implications)

• Methodological frameworks
  – Differentiate between design and appraisal/assessment

• Methodological challenges
  – Avoidance of risk of bias: e.g. Blinding/masking (lack of operator and patient blinding)
  – Heterogeneity: sensitivity analysis on the basis of intervention and comparator difference
  – Directness (applicability): judgments about directness challenging in relation to complex interventions

➡️ No reasons to accept different standards for the assessment of evidence from NPNTI, special measures to be taken to avoid risk of bias
Non-pharmaceutical, non-technical interventions (NPNTI)

- Erläutern von "Methodische Besonderheiten, Rahmenbedingungen und Herausforderungen bei Studien mit nicht-medikamentösen und nicht-technischen Behandlungsverfahren"
- Aim:
  - To clarify that this field, in terms of design and appraisal of research, suffers from similar problems as straightforward medical interventions
  - To suggest a framework that can be used to make comparisons against other fields in terms of our confidence in results

Example: respiratory rehabilitation (multi-component)
- Exercise
- Education
- Emotional support (psychological)
  - Smoking cessation

By definition: complex intervention (all interventions can be complex)
  - but not all complex interventions non-pharmaceutical
Avoidance of distorted effects

For the question (PICO) of interest

• Design

• Appraisal/assessment
Should patients with **recent** exacerbation of COPD undergo respiratory rehabilitation?

| Man 2004 |
|------------------|------------------|
| **Methods** | Randomised parallel group trial |
| **Participants** | 42 COPD patients (mean age 70 years, 41% males, FEV$_1$ =39% predicted) after inpatient treatment for acute exacerbation |
| **Interventions** | **Rehabilitation**: Multidisciplinary outpatient pulmonary rehabilitation (within 10 days of discharge) with endurance and strength exercise and patient education for 12 weeks (2 sessions/week). Completion rate of pulmonary rehabilitation of 85.7% (18 out of 21 patients)  
**Usual care**: Standard community care with respirologist. Follow-up: 12 weeks |

Design and execution features

Risk of bias

– sequence generation (selection bias)
– allocation sequence concealment (selection bias)
– blinding of participants and personnel (performance bias)
– blinding of outcome assessment (detection bias)
– incomplete outcome data (attrition bias)
– selective outcome reporting (reporting bias)
Confidence in estimates of effect

Bradford Hill Criteria
- Strength
- Consistency
- Temporality
- Biological gradient
- Specificity
- Biological Plausibility
- Coherence
- Experiment
- Analogy

Good, but insufficient (publication bias?)
Determinants of confidence: GRADE

- Any evidence
- 5 factors that can lower confidence
  1. limitations in detailed study design and execution (risk of bias criteria)
  2. Inconsistency (or heterogeneity)
  3. Indirectness (PICO and applicability)
  4. Imprecision
  5. Publication bias
- 4 factors can increase confidence
  1. Randomization
  2. large magnitude of effect
  3. opposing plausible residual bias or confounding
  4. dose-response gradient
Confidence in estimates of effect: GRADE

100% confident →

← GRADE’s starting point

0% confident →
Relation between PICO and available evidence
Relation between PICO and available evidence
Relation between PICO and available evidence
Assessing Quality of Evidence by Outcome

Table: GRADE’s approach to rating quality of evidence (aka confidence in effect estimates)
For each outcome based on a systematic review and across outcomes (lowest quality across the outcomes critical for decision making)

<table>
<thead>
<tr>
<th>Study design</th>
<th>Initial confidence in an estimate of effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized trials</td>
<td>High confidence</td>
</tr>
<tr>
<td>Observational studies</td>
<td>Low confidence</td>
</tr>
</tbody>
</table>

1. Establish initial level of confidence

2. Consider lowering or raising level of confidence

- Lower if
  - Risk of Bias
  - Inconsistency
  - Indirectness
  - Imprecision
  - Publication bias

- Higher if*
  - Large effect
  - Dose response
  - All plausible confounding & bias
    - would reduce a demonstrated effect or
    - would suggest a spurious effect if no effect was observed

3. Final level of confidence rating

- High
- Moderate
- Low
- Very low

*upgrading criteria are usually applicable to observational studies only.
Lowering confidence in RCTs

Table: GRADE's approach to rating quality of evidence (aka confidence in effect estimates)

For each outcome based on a systematic review and across outcomes (lowest quality across the outcomes critical for decision making)

1. Establish initial level of confidence

<table>
<thead>
<tr>
<th>Study design</th>
<th>Initial confidence in an estimate of effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized trials ➔</td>
<td>High confidence</td>
</tr>
<tr>
<td>Observational studies</td>
<td>Low confidence</td>
</tr>
</tbody>
</table>

2. Consider lowering or raising level of confidence

<table>
<thead>
<tr>
<th>Reasons for considering lowering or raising confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>↓ Lower if</td>
</tr>
<tr>
<td>Risk of Bias</td>
</tr>
<tr>
<td>Inconsistency</td>
</tr>
<tr>
<td>Indirectness</td>
</tr>
<tr>
<td>Imprecision</td>
</tr>
<tr>
<td>Publication bias</td>
</tr>
<tr>
<td>↑ Higher if*</td>
</tr>
<tr>
<td>Large effect</td>
</tr>
<tr>
<td>Dose response</td>
</tr>
<tr>
<td>All plausible confounding &amp; bias</td>
</tr>
<tr>
<td>• would reduce a demonstrated effect or</td>
</tr>
<tr>
<td>• would suggest a spurious effect if no effect was observed</td>
</tr>
</tbody>
</table>

3. Final level of confidence rating

<table>
<thead>
<tr>
<th>Confidence in an estimate of effect across those considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
</tr>
<tr>
<td>Moderate</td>
</tr>
<tr>
<td>Low</td>
</tr>
<tr>
<td>Very low</td>
</tr>
</tbody>
</table>

*upgrading criteria are usually applicable to observational studies only.
### Altering confidence in observational studies

**Table: GRADE's approach to rating quality of evidence (aka confidence in effect estimates)**

*For each outcome based on a systematic review and across outcomes (lowest quality across the outcomes critical for decision making)*

<table>
<thead>
<tr>
<th>Study design</th>
<th>Initial confidence in an estimate of effect</th>
<th>Reasons for considering lowering or raising confidence</th>
<th>Confidence in an estimate of effect across those considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized trials</td>
<td>High confidence</td>
<td>↓ Lower if</td>
<td>High (+++++)</td>
</tr>
<tr>
<td>Observational studies</td>
<td>Low confidence</td>
<td>↑ Higher if*</td>
<td>Moderate (++++)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Risk of Bias, Inconsistency, Indirectness, Imprecision, Publication bias</td>
<td>Low (++++), if applicable criteria</td>
</tr>
</tbody>
</table>

*upgrading criteria are usually applicable to observational studies only.*
Avoidance of distorted effects

For the question (PICO) of interest

• Design
### Design vs results

2. Consider lowering or raising level of confidence

<table>
<thead>
<tr>
<th>Reasons for considering lowering or raising confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lower if</strong></td>
</tr>
<tr>
<td>Risk of Bias</td>
</tr>
<tr>
<td>Inconsistency</td>
</tr>
<tr>
<td>Indirectness</td>
</tr>
<tr>
<td>Imprecision</td>
</tr>
<tr>
<td>Publication bias</td>
</tr>
</tbody>
</table>

Can be influenced by design (in addition to randomization)

Influenced by results
Reducing risk of bias in NTNPI

• Performance and detection bias (blinding)
• Choice of proper comparison
  – Evaluating effects of single/few components of an intervention (dismantling design)
• Expertise based RCTs
• Cluster RCTs
Reducing *risk of bias* in NTNPI

- **Respiratory Rehabilitation**
  - Main outcomes: Death, Hospitalization, HRQL
  - Impossible to blind
- **Challenge: Blinding/masking**
- **Depends on *real life* PICO**
  - When comparator “usual care” lack of blinding bigger problem than when active comparator
- **Intervention and effect?**
  - Package vs individual components
Reducing *risk of bias* in NTNPI

It should be recognized that all placebos in psychotherapy contain one fundamental flaw: In the placebo condition therapists are required allegiance to that treatment. That allegiance is an important determinant of outcome in psychotherapy (Wampold, 2001b) suggests that the therapists may be less enthusiastic, less hopeful, less engaged, or less empathic in such conditions; that is, the

Design & conduct *dismantling studies (RCTs)*

A + B + C + D vs A + B + C

“Straightforward”
Interval versus Continuous High-Intensity Exercise in Chronic Obstructive Pulmonary Disease

A Randomized Trial

Milo A. Puhan, MD, PhD; Gilbert Büsching, PT; Holger J. Schünemann, MD, MSc, PhD; Evelien vanOort, PT; Christian Zaugg, PhD; and Martin Frey, MD

Background: Guidelines recommend high-intensity continuous exercise to reduce peripheral muscle dysfunction in patients with chronic obstructive pulmonary disease but acknowledge that interval exercise might be an equally effective alternative that is better tolerated by patients.

Objective: To assess whether interval exercise is no less effective than high-intensity continuous exercise and whether it is tolerated better by patients with severe chronic obstructive pulmonary disease.

Design: Randomized, noninferiority trial.

A + B + C + D vs A + B + C + E
Those administering the intervention rarely have equal skills in (both) alternatives.
Participant vs expertise-based RCTs

• In regular RCTs, investigators need to ensure that number of those with expertise in the alternatives is equal.
• If one of the two alternatives is more challenging to perform and requires more training (operator experience), then there may be effect overestimation of the less challenging procedure, because those unfamiliar with the more challenging procedure may be unable to perform at a high level.
Expertise based RCTs

• Randomization of patients to operators with expertise in the intervention
  – Operators likely biased towards one procedure
  – Differential procedure bias will decrease
  – Outcome assessment bias decreases
  – Procedural cross-over decreases because operators “get to do what they want”
Avoidance of distorted effects

For the question (PICO) of interest

• Design

• Appraisal/assessment
Systematic reviews of Respiratory Rehabilitation studies

pulmonary rehabilitation compared to usual community care for COPD with recent exacerbation

Patient or population: patients with COPD with recent exacerbation
Settings: outpatient
Intervention: pulmonary rehabilitation
Comparison: usual community care

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Illustrative comparative risks* (95% CI)</th>
<th>Relative effect (95% CI)</th>
<th>No of Participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assumed risk</td>
<td>Corresponding risk</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Usual community care</td>
<td>Pulmonary rehabilitation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital admission</td>
<td>405 per 1000</td>
<td>130 per 1000 (52 to 283)</td>
<td>OR 0.22 (0.08 to 0.58)</td>
<td>250</td>
<td>⬤⬤⬤⬤ high</td>
</tr>
<tr>
<td>Follow-up: 3-18 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality</td>
<td>Low</td>
<td>100 per 1000 (11 to 85)</td>
<td>OR 0.28 (0.1 to 0.84)</td>
<td>110</td>
<td>⬤⬤⬤ moderate²</td>
</tr>
<tr>
<td>Follow-up: 3-48 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>500 per 1000 (91 to 457)</td>
<td></td>
<td>258</td>
<td>⬤⬤⬤ moderate⁴</td>
</tr>
<tr>
<td>Quality of life (CRQ) dyspnea</td>
<td>Chronic Respiratory Questionnaire³. Scale from: 1 to 7.</td>
<td>The mean quality of life (crq) dyspnea in the intervention groups was 0.37 higher (0.35 to 1.58 higher)</td>
<td>258</td>
<td>⬤⬤⬤ moderate⁴</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Follow-up: 12 and 76 weeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quality of life (SGRQ) total</td>
<td>St George's Respiratory Questionnaire². Scale from: 0 to 100.</td>
<td>The mean quality of life (sgrq) total in the control groups was 9.88 lower (5.37 to 14.4 lower)</td>
<td>127</td>
<td>⬤⬤⬤ moderate⁴</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Follow-up: 12 and 26 weeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambulation (as measured by 6 min walking distance)</td>
<td>distance in meters⁶</td>
<td>The mean ambulation (as measured by 6 min walking distance) in the intervention groups was 77.7 higher (12.21 to 143.2 higher)</td>
<td>299</td>
<td>⬤⬤⬤ moderate⁴, ⁸</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Follow-up: 1 - 208 weeks³</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resource use - not reported</td>
<td>See comment</td>
<td>See comment</td>
<td>Not estimable</td>
<td>-</td>
<td>See comment</td>
</tr>
</tbody>
</table>

¹ Median of baseline risk in studies used for low risk estimate. High risk estimate of 50% based on study with 4 years of follow-up.
² Only 23 events total.
³ CRQ: MID 0.5, moderate effect 1.0, large effect 1.5
⁴ Sample size substantially lower than 400
⁵ SGRQ: MID: 4, moderate effect 8, large effect 12.
⁶ 6MWD: MID is 35 m (30 to 42) or 10% change of baseline 6 min walk test
⁷ 76 weeks, 11 days, 6 weeks, 208 weeks, in the 4 trials respectively
⁸ Although I² square is 89%, this significant heterogeneity is likely due to large differences in baseline severity (baseline 6 min walk test distance)
Figure 8. Forest plot of comparison: 1 Rehabilitation versus control, outcome: 1.5 Change from baseline in 6 minute walking test.

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Mean Difference</th>
<th>SE</th>
<th>Weight</th>
<th>Mean Difference IV, Random, 95% CI</th>
<th>Mean Difference IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behnke 2000</td>
<td>215</td>
<td>28</td>
<td>16.0%</td>
<td>215.00 [160.12, 269.88]</td>
<td></td>
</tr>
<tr>
<td>Carr 2009</td>
<td>-25</td>
<td>23</td>
<td>16.6%</td>
<td>-25.00 [-70.08, 20.08]</td>
<td></td>
</tr>
<tr>
<td>Eaton 2009</td>
<td>-2</td>
<td>16</td>
<td>17.4%</td>
<td>-2.00 [-33.36, 29.36]</td>
<td></td>
</tr>
<tr>
<td>Kirsten 1998</td>
<td>158</td>
<td>28</td>
<td>16.0%</td>
<td>158.00 [103.12, 212.88]</td>
<td></td>
</tr>
<tr>
<td>Nava 1998</td>
<td>68</td>
<td>19</td>
<td>17.1%</td>
<td>68.00 [30.76, 105.24]</td>
<td></td>
</tr>
<tr>
<td>Troosters 2000</td>
<td>64</td>
<td>21</td>
<td>16.9%</td>
<td>64.00 [22.84, 105.16]</td>
<td></td>
</tr>
<tr>
<td><strong>Total (95% CI)</strong></td>
<td><strong>100.0%</strong></td>
<td></td>
<td></td>
<td><strong>77.70 [12.21, 143.20]</strong></td>
<td></td>
</tr>
</tbody>
</table>

Heterogeneity: $\text{Tau}^2 = 6179.55$; $\text{Chi}^2 = 71.60$, df = 5 ($P < 0.00001$); $I^2 = 93\%$

Test for overall effect: $Z = 2.33$ ($P = 0.02$)
### 1.3.3 CRQ: Emotional function domain

<table>
<thead>
<tr>
<th>Study</th>
<th>Effect Size</th>
<th>Standard Error</th>
<th>Percentage</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behnke 2000</td>
<td>1.78</td>
<td>0.45</td>
<td>15.5%</td>
<td>1.78</td>
<td>[0.90, 2.66]</td>
</tr>
<tr>
<td>Carr 2009</td>
<td>0.70</td>
<td>0.40</td>
<td>17.4%</td>
<td>0.70</td>
<td>[-0.08, 1.48]</td>
</tr>
<tr>
<td>Eaton 2009</td>
<td>0.40</td>
<td>0.21</td>
<td>26.1%</td>
<td>0.40</td>
<td>[-0.01, 0.81]</td>
</tr>
<tr>
<td>Man 2004</td>
<td>1.24</td>
<td>0.24</td>
<td>24.7%</td>
<td>1.24</td>
<td>[0.77, 1.71]</td>
</tr>
<tr>
<td>Seymour 2010</td>
<td>0.80</td>
<td>0.43</td>
<td>16.3%</td>
<td>0.80</td>
<td>[-0.04, 1.64]</td>
</tr>
<tr>
<td><strong>Subtotal (95% CI)</strong></td>
<td><strong>100.0%</strong></td>
<td><strong>0.94</strong></td>
<td><strong>0.46, 1.42</strong></td>
<td><strong>100.0%</strong></td>
<td><strong>0.46, 1.42</strong></td>
</tr>
</tbody>
</table>

Heterogeneity: Tau² = 0.19; Chi² = 11.66, df = 4 (P = 0.02); I² = 66%
Test for overall effect: Z = 3.81 (P = 0.0001)
Indirectness - intervention

Outpatient respiratory rehabilitation including exercise, education and emotional support (in patients with COPD)

Respiratory rehabilitation Inpatient program

No concerns about directness (transferability)
No lowering of confidence
Same recommendation

Concerns about directness
Lower confidence
Separate recommendation
Indirectness - intervention

Outpatient respiratory rehabilitation including exercise, education and emotional support (in patients with COPD)

Respiratory rehabilitation inpatient programs

Similarities in:
Exercise program
Education
Emotional support
Co-interventions?
And how large is the contribution of each component

No concerns about directness (transferability)
No lowering of confidence
Same recommendation

Concerns about directness
Lower confidence
Separate recommendation
## Directness?

<table>
<thead>
<tr>
<th>Domain (original question asked)</th>
<th>Description (evidence found and included, including evidence from other studies) – consider the domains of study design and study execution, inconsistency, imprecision and publication bias</th>
<th>Judgment - Is the evidence sufficiently direct?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population:</strong></td>
<td></td>
<td>Yes □  Probably yes □  Probably no □  No □</td>
</tr>
<tr>
<td><strong>Intervention:</strong></td>
<td></td>
<td>Yes □  Probably yes □  Probably no □  No □</td>
</tr>
<tr>
<td><strong>Comparator:</strong></td>
<td></td>
<td>Yes □  Probably yes □  Probably no □  No □</td>
</tr>
<tr>
<td><strong>Direct comparison:</strong></td>
<td></td>
<td>Yes □  Probably yes □  Probably no □  No □</td>
</tr>
<tr>
<td><strong>Outcome: Mortality</strong></td>
<td></td>
<td>Yes □  Probably yes □  Probably no □  No □</td>
</tr>
<tr>
<td><strong>Final judgment about indirectness across domains:</strong></td>
<td>□ No indirectness □ Serious indirectness □ Very serious indirectness</td>
<td></td>
</tr>
</tbody>
</table>

Schünemann, et al. Systematic reviews (in press)
Assessing Quality of Evidence by Outcome

**Table: GRADE’s approach to rating quality of evidence (aka confidence in effect estimates)**

*For each outcome based on a systematic review and across outcomes (lowest quality across the outcomes critical for decision making)*

<table>
<thead>
<tr>
<th>Study design</th>
<th>Initial confidence in an estimate of effect</th>
<th>Reasons for considering lowering or raising confidence</th>
<th>Confidence in an estimate of effect across those considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized trials ➔</td>
<td>High confidence</td>
<td>Lower if: Risk of Bias, Inconsistency, Indirectness, Imprecision, Publication bias</td>
<td>High (+++++)</td>
</tr>
<tr>
<td>Observational studies ➔</td>
<td>Low confidence</td>
<td>Higher if*: Large effect, Dose response, All plausible confounding &amp; bias (• would reduce a demonstrated effect or • would suggest a spurious effect if no effect was observed)</td>
<td>Moderate (++++)</td>
</tr>
</tbody>
</table>

*upgrading criteria are usually applicable to observational studies only.*
Design vs results

2. Consider lowering or raising level of confidence

<table>
<thead>
<tr>
<th>Reasons for considering lowering or raising confidence</th>
<th>↓ Lower if</th>
<th>↑ Higher if</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk of Bias</td>
<td>Large effect</td>
<td></td>
</tr>
<tr>
<td>Inconsistency</td>
<td>Dose response</td>
<td></td>
</tr>
<tr>
<td>Indirectness</td>
<td>All plausible confounding &amp; bias</td>
<td></td>
</tr>
<tr>
<td>Imprecision</td>
<td>• would reduce a demonstrated effect or</td>
<td></td>
</tr>
<tr>
<td>Publication bias</td>
<td>• would suggest a spurious effect if no effect was observed</td>
<td></td>
</tr>
</tbody>
</table>

Can be influenced by design (in addition to randomization)

Influenced by results
Those designing research should use structured framework to determine which measures they can take to enhance overall confidence in results they obtain.
- Separate design from results based confidence
- Heterogeneity: sensitivity analysis on the basis of intervention and comparator difference
- Directness (applicability): judgments required, no placebo or control is perfect

Those appraising research can use same structured approach but will place greater emphasis on all factors.

No reason to accept different standards for the assessment of the evidence related to NPNTI
THANK YOU