Forskningsdesigns og evidens
Aims of the presentation

• Evaluating the available evidence – challenges and available solutions
  – Problems in every-day thinking
  – Aims of scientific method
  – Types of evidence
  – Threats to validity
  – When one study is not enough
  – Aspects to consider when evaluating “effectiveness”
Unit for Psychooncology and Health Psychology - EPoS

- Psychooncology Research Unit established in 2000 at the Dept. Of Oncology, AUH based on a grant from the Danish Cancer Society
- EPoS established in 2011 in collaboration between AUH, Dept. Of Oncology, BSS, AU, and Dept of Psychology and Behavioural Science
- Current staff: 17 (1 professor, 2 assoc. prof. 1 assist prof, 1 senior researcher, 4 post-docs, 7 PhD’s, 1 adm.) + 8-10 research assistants.
My background

• Research areas
  – Psychoneuroimmunology
  – Pain research
  – Psycho-social cancer research
  – Patient-health professional interactions
  – Health psychology
  – Psychosocial interventions
  – Hypnosis, guided imagery, mindfulness-based intervention, expressive writing
  – Internet-delivered interventions
  – Efficacy of complementary and alternative treatments
My background

• Research methodologies
  – Experimental studies (psychophysiology)
  – Randomized Controlled Trials (RCT)
  – Pragmatic trials
  – Mixed (qualitative and quantitative) methods
  – Cross-sectional and cohort studies
  – Psychometrics
  – Systematic reviews and meta-analysis
• **Intellectual advise to future generations:**

• “Ask yourself only what are the facts and what is the truth that the facts bear out. Never let yourself be diverted either by what you wish to believe, or by what you think would have beneficent social effects if it were believed. But look only, and solely, at what are the facts.”

Bertrand Russell, 1872-1970
Why scientific method?

*Personal experience vs Scientific evidence*
"Man prefers to believe what he prefers to be true"
Automatic vs manual processing
“Fast” vs “slow” thinking

System 1
Fast
Automatic
Emotional
Frequent
Stereotypic
Subconscious

System 2
Slow
Effortful
Logical
Infrequent
Calculating
Conscious

Evolution has equipped humans with two types of thinking

Shiv et al. 2005; Kahneman, 2011

Automatic mode
Ventromedial prefrontal cortex

Manual mode
Dorsolateral prefrontal cortex
Problems in Everyday- (fast) Thinking

• The practical value of everyday ("fast") thinking is obtained at the cost of bias and precision, e.g.,:
  – We tend to see patterns, even where there are none
  – We see causal relationships, even where there are none
  – We tend to focus on and remember positive evidence
  – We tend to overestimate evidence confirming our position
  – Our judgments are influenced by the judgments of our surroundings (conformity)
  – We tend to believe that positive and negative traits, respectively, are associated (clustering)
  – We tend to overestimate the probability of dramatic events
Problems in fast thinking

• The practical value of everyday ("fast") thinking is obtained at the cost of bias and precision, e.g.,
  – We have a tendency to see patterns, also when there are no patterns:
    • E.g. in completely random sequences: 1222122212211121112211.
    • E.g. When rolling dice: If we have not obtained a "six" in many rolls, we tend to believe that the probability increases (although the chance remains 1/6 even after 100 rolls)
  – We tend to perceive causal relationships, also when there are none:
    • E.g., we tend to imagine causality between X and Y, if Y takes place after X (a necessary but not sufficient condition for causality)
The echo chamber problem

- Closed ideology echo chamber
- Applies to politics as well as other domains

The solution is a no-brainer – but difficult to practice
“We love to predict things – and we aren’t very good at it”

Nate Silver (1978 -)
The domains of science

- **Meta-physics**: Philosophy, epistemology: deduction and reasoning
- **Theories of science**: Meta-theories about method
- **Theoretical science**: Collecting, condensing, discussing, and interpreting existing theoretical and empirical research results
- **Empirical science**: Measuring phenomena and testing hypotheses
  - Observing and describing
  - Predicting
  - Determining causes
  - Explaining
The aim of scientific method

• **General aim**: To generate measurable and testable data, gradually adding to the accumulation of human knowledge
  – To produce *reliable* knowledge
  – To produce *valid* knowledge
  – About *causal* relationships
  – By addressing possible *sources of error*
The falsification principle

Karl Popper: A Scientific Hypothesis Must Be "Falsifiable".

- We support a hypothesis by falsifying the null-hypothesis
- A general approach: We do not “prove” hypotheses – but maximize our attempts to falsify statements about observations, associations, causality, and mechanisms
Types of evidence
Evidence hierarchy

Primary research

Clinical experience

Experimental design (RCT)
Cohort study
Case-control study
Cross sectional study
Case reports
Expert opinion
Anecdote

Challenge: the model favors internal validity
Types of validity

- Validity
  - Internal validity (causality, excluding alternative explanations, sources of error)
  - External validity (generalizability)
  - Ecological validity (pragmatic validity)

- The three types of validity supplement each other
- Are difficult to obtain with one single method
- Internal validity a prerequisite for external and ecological validity
- Reliability a prerequisite for validity – but not the reverse
Mistaking reliability for validity

Accuracy of matching charts to owners N = 54

Approx 50% confidence limits

Can clients pick own reading? N = 18

Agreement between astrologers N = 27
Research questions

• Internal validity
  – Does it work (statistical significance, superiority)?
  – How well does it work (practical significance)?
  – Does it work as well as something else (non-inferiority)?
  – How does it work (mechanisms, specific, non-specific)?

• External validity
  – For whom does it work?
  – For whom does it not work?

• Ecological validity
  – Does it work in the clinical practical context?
  – Clinician and patient adherence
Example of design

*maximizing internal validity*: Randomized controlled trial
MBCT for persistent pain in women treated for breast cancer

- 16-20% of women treated for breast cancer experience pain after 5-9 years
- Limited pharmacological treatment efficacy
- Pain is a multidimensional phenomenon consisting of sensory, cognitive, and affective factors
- Mindfulness-based therapy teaches ways of relating to bodily sensations and emotional discomfort with higher degree of acceptance and openness
- Mindfulness-Based Cognitive Therapy may be effective for cancer-related pain
MBCT for persistent pain in women treated for breast cancer

Efficacy of Mindfulness-Based Cognitive Therapy on Late Post-Treatment Pain in Women Treated for Primary Breast Cancer: A Randomized Controlled Trial

Purpose
To assess the efficacy of mindfulness-based cognitive therapy (MBCT) for late post-treatment pain in women treated for primary breast cancer.

Methods
A randomized wait-list controlled trial was conducted with 120 women treated for breast cancer reporting post-treatment pain score ≥3 or pain intensity, or pain burden assessed with SF-36 and anxiety and depression scales. Participants were randomly assigned to a wait-list control group or a 12-week MBCT program or a wait-list control group. Pain was the primary outcome and was assessed with the Short Form McGill Pain Questionnaire (SF-MPQ-2), the Present Pain Intensity subscale of the McGill Pain Questionnaire, and mean pain intensity and pain burden ratings. Secondary outcomes were quality of life, anxiety and depression, and self-reported use of pain medication. All outcomes were measured at baseline, post-intervention, and 3, 6, and 12 months follow-up. Treatment effects were assessed using mixed model regression.

Results
Statistically significant time × group interactions were found for pain intensity (t = 0.61, F = 0.02), the Present Pain Intensity subscale (t = 0.50, F = 0.02), the SF-MPQ-2 neuropathic pain subscale (t = 0.50, F = 0.02), and SF-MPQ-2 total score (t = 0.50, F = 0.02). Only pain intensity remained statistically significant after correction for multiple comparisons. Statistically significant effects were also observed for quality of life (t = 0.42, F = 0.02) and non-opioid pain medications (t = 0.48, F = 0.02). None of the remaining outcomes reached statistical significance.

Conclusion
MBCT showed a statistically significant, robust, and durable effect on pain intensity, indicating that MBCT might be an effective pain rehabilitation strategy for women treated for breast cancer. In addition, the effect on neuropathic pain, a pain type reported for women treated for breast cancer, further suggests the potential of MBCT may be considered preliminary.

Zachariae
MBCT for persistent pain in women treated for breast cancer

- *Mediators* of the effect of MBCT on pain intensity:
  - Mindfulness non-reactivity facet; Pain catastrophizing

![Diagram showing the relationship between Group (MBCT versus control), Non-reactivity (1), Pain catastrophizing (2), and Pain intensity.]

Johannsen, O’Toole, O’Connor, Jensen & Zachariae (under review)
MBCT for persistent pain in women treated for breast cancer

Cost-effectiveness of MBCT for persistent pain in women treated for breast cancer

Johannsen, Sørensen, O’Connor, O’Toole, Zachariae (in preparation)
Example of design
Pragmatic trial maximizing both internal, external, and ecological validity
Evaluating complementary and alternative treatments

- Lab experiments and RCTs of energy healing yield negative results
  Zachariae et al. 2005

- Complementary and alternative medicines and treatments (CAMs) are associated with increased symptoms of depression in breast cancer patients
  Pedersen et al. 2013

- Active acupuncture is not more effective than placebo acupuncture in a double blinded RCT
  Vase et al. 2013
Effectiveness of energy healing on quality of life in colorectal cancer patients

• CAM practitioners criticize validity of traditional research methodologies, e.g., RCTs

• Criticisms include:
  – May not be generalized to the general population
  – People may have strong treatment preferences
  – Standardized outcome measures may not cover patients’ individual concerns
  – Patients may prefer some practitioners to others
  – Standardization of treatment context may cancel out effects
Effectiveness of energy healing on quality of life in colorectal cancer patients

- Study designed to maximize internal, external, and ecological validity
- Colorectal cancer patients randomized to:
  - A Randomization
    - Healing or control
  - B Self-selection
    - Healing or control
- Patients
  - Selected their healer from a list
  - Treatment took place in healer’s clinic
  - Completed standardized QoL measures
  - Prioritized preferred outcome
  - Completed measure of attitude towards CAM
Effectiveness of energy healing on quality of life in colorectal cancer patients

- No overall effects on any outcomes
- Small effect on QoL in subgroup: Patients in self-selected healing group who had rated QoL as important, and who had a positive attitude towards CAM
Example of design
Pragmatic evaluation of daily practice
Pragmatic evaluation

• Fagligt Selskab for Psykologer i Palliation og Onkologi
• 11 psychologists treating 92 patients or caregivers
• Psychologist questionnaires: psychotherapeutic models and tools used
• Patient questionnaires pre- and post consultation:
  – MYCaW (Measure Yourself Concerns and Well-being): Primary and secondary concern and general well-being
  – Working Alliance Inventory
• N of 1 statistics: Reliable Change Index (RCI)
  – Determines whether a change is beyond a statistical error

Pedersen et al. (unpublished)
Pragmatic evaluation

<table>
<thead>
<tr>
<th>Selvvalgt problemområde og alment velbefindende</th>
<th>Sign. (RCI) Forbedring</th>
<th>Ingen ændring</th>
<th>Sign. (RCI) Forværring</th>
<th>Data mangler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primært problemområde</td>
<td>N (%)</td>
<td>N (%)</td>
<td>N (%)</td>
<td>N (%)</td>
</tr>
<tr>
<td>22 (23,9)</td>
<td>64 (69,6)</td>
<td>0 (0,0)</td>
<td>6 (6,5)</td>
<td></td>
</tr>
<tr>
<td>Sekundært problemområde</td>
<td>20 (21,7)</td>
<td>43 (46,7)</td>
<td>0 (0,0)</td>
<td>29 (31,5)</td>
</tr>
<tr>
<td>Alment velbefindende</td>
<td>28 (30,4)</td>
<td>60 (65,2)</td>
<td>1 (1,1)</td>
<td>3 (3,3)</td>
</tr>
</tbody>
</table>

Statistically significant predictors of sign. improvement:
- Positive expectancies: The session will improve my understanding of my reactions and emotions
- Perceived working alliance
- Higher educational level

Pedersen et al. (unpublished)
Interpreting results

The Earth Is Round \( (p < .05) \)

Jacob Cohen

After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists. This article reviews the problems with this practice, including its near-universal misinterpretation of \( p \) as the probability that \( H_0 \) is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects \( H_0 \) one thereby affirms the theory that led to the test. Exploratory data analysis and the use of graphic methods, a steady improvement in and a movement toward standardization in measurement, an emphasis on estimating effect sizes using confidence intervals, and the informed use of available statistical methods is suggested. For generalization, psychologists must finally rely, as has been done in all the older sciences, on replication.

sure how to test \( H_0 \) chi-square with Yates’s (1951) correction or the Fisher exact test, and wonders whether he has enough power. Would you believe it? And would you believe that if he tried to publish this result without a significance test, one or more reviewers might complain? It could happen.

Almost a quarter of a century ago, a couple of sociologists, D. E. Morrison and R. E. Henkel (1970), edited a book entitled The Significance Test Controversy. Among the contributors were Bill Rozeboom (1960), Paul Meehl (1967), David Bakan (1966), and David Lykken (1968). Without exception, they damned NHST. For example, Meehl described NHST as "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" (p. 265). They were, however, by no means the first to do so. Joseph Berkson attacked NHST in 1938, even before it
What is an effect size?
A standardized effect, e.g., standardized mean difference, enabling comparisons across measures and studies

Cohen’s $d = (\text{Mean 1} - \text{Mean 2})/\text{SD (pooled)}$

Which intervention is most effective?

- **Study 1**: Mean score and standard deviation of Hamilton Depression Rating Scale (range: 0-49):
  - Intervention: 16.5 (13.0)
  - Control: 20.5 (14.0)
  - Cohen’s $d = \boxed{0.29}$

- **Study 2**: Mean score and standard deviation of Beck’s Depression Inventory (range: 0-63)
  - Intervention: 17.5 (7.0)
  - Control: 21.5 (8.0)
  - Cohen’s $d = \boxed{0.53}$

HDRS MCID (0.5 SD) \(^1\)

BDI MCID (17%) \(^2\) = 3.7 = SD: 0.49

*) To detect the difference in $d$ between study 1 and 2 requires a sample of 610 in each group

1) NICE, 2004; 2) Button et al. 2015
Significance and precision

Both effect sizes are statistically significantly different from "0", are not different from each other

Only the effect size of study 1 is significantly different from "0". The two effect sizes are not different from each other

Both effect sizes are statistically significantly different from "0", and two effect sizes are sign. different from each other ($p = 0.003$)

Significance: $p$-values < 0.05; Precision: 95% Confidence interval
One study is not enough!

Challenges:
Non-replication publication bias
“cherry picking”
Interpreting non-replicated results

Why Most Published Research Findings Are False
John P. A. Ioannidis

Summary
There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller, when effects are measured on a greater number and lesser prescreening of tests/relationships, where there is greater flexibility in design, definitions, outcomes, and analytical models, when there is greater financial and other interest and competition, and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for any study design and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, published research findings may often be simply inaccurate measurements of the prevailing bias. This work elucidates the implications of these possibilities for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with causing confusion and disappointment. Replication and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1-7] to the most modern molecular research [8-11]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research results [12-14]. The probability that a research finding is true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance. Considering a 2 × 2 table in which research findings are compared across two studies, and some studies yield positive results, while others yield negative results, the positive predictive value (PPV) is given by the following formula:

PPV = (TP / (TP + FP))

Where TP is the true positive and FP is the false positive.

C Onclusions
Contradiction and initially stronger effects are not unusual in highly cited clinical research and their outcomes. Extent to which high citations may provide confounders and also needs more study. Controversies are most common with highly cited, nonrandomized studies, but even the most highly cited randomized trials may be challenged and refuted over time.

Reference

Copyright © 2005 John P. A. Ioannidis. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviations
PPV: positive predictive value

John P. A. Ioannidis is in the Department of Biostatistics and Epidemiology, University at Buffalo School of Medicine and Biomedical Sciences, Buffalo, New York 14214, USA.

Zachariae
“Replication crisis”

Is treatment X better than control?  
(Note: A smaller value is better)

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th></th>
<th></th>
<th>Control</th>
<th></th>
<th></th>
<th></th>
<th>Sign.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anderson et al. 2000</td>
<td>60,3</td>
<td>18</td>
<td>13</td>
<td>68,9</td>
<td>16,7</td>
<td>12</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>Burrows et al. 2001</td>
<td>57,9</td>
<td>11,9</td>
<td>20</td>
<td>63,7</td>
<td>11,3</td>
<td>19</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>Crapper et al. 2002</td>
<td>66,5</td>
<td>19,5</td>
<td>14</td>
<td>59,8</td>
<td>22,9</td>
<td>13</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>Dim et al. 2003</td>
<td>56,9</td>
<td>16,7</td>
<td>17</td>
<td>79,1</td>
<td>17,2</td>
<td>17</td>
<td>p &lt; 0.01</td>
<td></td>
</tr>
<tr>
<td>Epstein et al. 2004</td>
<td>61,4</td>
<td>17,9</td>
<td>14</td>
<td>65,9</td>
<td>16,8</td>
<td>14</td>
<td>ns</td>
<td></td>
</tr>
<tr>
<td>Fleizig et al. 2005</td>
<td>58</td>
<td>11,9</td>
<td>15</td>
<td>61,5</td>
<td>11,3</td>
<td>15</td>
<td>ns</td>
<td></td>
</tr>
</tbody>
</table>
Evidence hierarchy

- Secondary research
- Primary research
- Clinical experience

Quantitative systematic review
Narrative systematic review
Experimental design (RCT)
Cohort study
Case-control study
Cross sectional study
Case reports
Expert opinion
Anecdote

Challenge: the model favors internal validity
Systematic reviews and meta-analysis

Single studies

Systematic review and meta-analysis
Results of meta-analysis:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Control</th>
<th>Sample size</th>
<th>Heterogeneity</th>
<th>Global effect sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>K</td>
<td>N</td>
<td>Q</td>
</tr>
<tr>
<td>Anderson et al. 2000</td>
<td>60.3</td>
<td>18</td>
<td>13</td>
<td>68.9</td>
</tr>
<tr>
<td>Burrows et al. 2001</td>
<td>57.9</td>
<td>11.9</td>
<td>20</td>
<td>63.7</td>
</tr>
<tr>
<td>Crapper et al. 2002</td>
<td>66.5</td>
<td>19.5</td>
<td>14</td>
<td>59.8</td>
</tr>
<tr>
<td>Dim et al. 2003</td>
<td>56.9</td>
<td>16.7</td>
<td>17</td>
<td>79.1</td>
</tr>
<tr>
<td>Epstein et al. 2004</td>
<td>61.4</td>
<td>17.9</td>
<td>14</td>
<td>65.9</td>
</tr>
<tr>
<td>Fleizig et al. 2005</td>
<td>58</td>
<td>11.9</td>
<td>15</td>
<td>61.5</td>
</tr>
</tbody>
</table>

Zachariae
Systematic reviews and meta-analysis

- Can test (falsify) hypotheses
- Test reproducibility (were initial results random or reliable?)
- Control for random error (variation) between studies
- Test systematic variation between studies
- Generalize results
Number of published meta-analyses per year - PubMED 1990-2013
History of meta-analysis

- **Pearson (1904)** averaged correlations between mortality and inoculation for typhoid fever.
- First medical “meta-analysis” on placebo effects (**Beecher, 1955**)
- **Eysenck (1952)** argued that psychotherapy was ineffective
- Glass standardized and averaged treatment-control differences from 375 studies, naming it ”meta-analysis” (**Smith & Glass, 1977**)
- ”An exercise in mega-silliness” (**Eysenck, 1978**)
- Similar methods developed by **Rosenthal and Rubin (1978)**
- **Cochrane Collaboration (1993)**: Medicine
- **Campbell Collaboration (1999)**: Social sciences
- **Handbook of research synthesis** (**Cooper & Hedges, 1994**)
Narrative vs Systematic review: A matter of life or death

- From 1972-81, 7 studies investigated the effect of steroid-injektions on premature delivery (associated with increased infant mortality)
- Two studies showed a weak positive effect – the remaining studies were non-significant
- The treatment was abandoned
- A later 1989 meta-analysis of the original data revealed a significant positive effect on infant mortality (OR: 0.50)
- The Cochrane Collaboration logo shows data from the 1989 meta-analysis

www.cochrane.org
Risk of bias

“Excellent health statistics - smokers are less likely to die of age related illness”
Study quality

• **Validity**: “The approximate truth of an inference or claim about a relationship”

• **Internal validity**
  – Threats: all alternative mechanisms that could explain results, e.g., “placebo”, group-differences at baseline, uneven dropout

• **External validity**
  – Are results generalizable to other intended participants and contexts?

• **Construct validity**
  – Do the operational characteristics of intervention and measures adequately represent intended abstract categories?

• **Statistical conclusion validity**
  – The validity of the statistical inferences regarding the strength of the relationship. Threats include insufficient statistical power, regression towards the mean, incorrect assumptions about the underlying variance

Shadish, Cook & Campbell, 2002
Quality assessment

- All studies should be subjected to a predefined quality assessment
- Already developed or modified existing checklist
- A newly developed checklist

E.g., Jadad checklist (Jadad, 1996)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>score (1, 0, or −1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1. Did the study include and report both a pre- and post-assessment? ***</td>
<td></td>
</tr>
<tr>
<td>C2. Was the study randomized?</td>
<td></td>
</tr>
<tr>
<td>C3. Was the method of randomization described and appropriate?</td>
<td></td>
</tr>
<tr>
<td>C4. Was the randomization described BUT inappropriate? (deduct 1 pt)</td>
<td></td>
</tr>
<tr>
<td>C5. Was the study described as double blind?</td>
<td></td>
</tr>
<tr>
<td>C6. Was any form of blinding of condition (patients) attempted?</td>
<td></td>
</tr>
<tr>
<td>C7. Was any form of blinding of assessment/results (researchers) attempted?</td>
<td></td>
</tr>
<tr>
<td>C8. Was the blinding described and appropriate (e.g. attention control)?</td>
<td></td>
</tr>
<tr>
<td>C9. Was blinding described but inappropriate? (deduct 1 pt)</td>
<td></td>
</tr>
<tr>
<td>C10. Was there a description of withdrawals and dropouts?</td>
<td></td>
</tr>
<tr>
<td>C11. Were the objectives of the study defined?***</td>
<td></td>
</tr>
<tr>
<td>C12. Were the outcome measures defined clearly?</td>
<td></td>
</tr>
<tr>
<td>C13. Was there a clear description of the inclusion and exclusion criteria?</td>
<td></td>
</tr>
<tr>
<td>C14. Was the sample size justified (e.g. power calculation)?</td>
<td></td>
</tr>
<tr>
<td>C15. Was there a clear description of the intervention?</td>
<td></td>
</tr>
<tr>
<td>C16. Was there at least one control (comparison group)?</td>
<td></td>
</tr>
<tr>
<td>C17. Was the method used to assess adverse effects described?</td>
<td></td>
</tr>
<tr>
<td>C18. Were the methods of statistical analysis described?</td>
<td></td>
</tr>
<tr>
<td>C19. Were the outcome measure used standardized and reliable?</td>
<td></td>
</tr>
<tr>
<td>Jadad 3-item score total (range: 0 – 5)</td>
<td></td>
</tr>
<tr>
<td>Jadad 11-item score total (range: 0 – 13)</td>
<td></td>
</tr>
<tr>
<td>Total (modified) quality score (range: 0 – 17)</td>
<td></td>
</tr>
</tbody>
</table>

Zachariae
### Quality assessment

#### Masking conditions

- Randomized
- Clear description of randomization
- Attempts to mask condition to part.
- Concealment of allocation to res.
- Dropouts clearly described
- Objectives clearly described
- Outcome measures clearly described
- Inclusion-exclusion clearly described
- Sample size justified e.g. power analysis
- Intervention clearly described
- Control condition included
- Statistics clearly described
- Free of suggestions of selective outcome rep.
- Manipulation check included and described
- Active control condition included

#### Power analysis

- Study
- Randomized
- Clear description of randomization
- Attempts to mask condition to part.
- Concealment of allocation to res.
- Dropouts clearly described
- Objectives clearly described
- Outcome measures clearly described
- Inclusion-exclusion clearly described
- Sample size justified e.g. power analysis
- Intervention clearly described
- Control condition included
- Statistics clearly described
- Free of suggestions of selective outcome rep.
- Manipulation check included and described
- Active control condition included

- Possible score: 0-15
- Mean score = 11.3 (SD = 2.4; range: 8-15)
- Inter-rater agreement: 89.9% of 225 individual quality ratings

---

Zachariae & O’Toole, 2015
ES of published vs unpublished

Table 3
Comparison of Effect Sizes Reported in Published Versus Unpublished Studies

<table>
<thead>
<tr>
<th>Document source</th>
<th>Effect size</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published studies</td>
<td>0.53</td>
<td>0.30</td>
<td>92</td>
</tr>
<tr>
<td>Unpublished studies</td>
<td>0.39</td>
<td>0.28</td>
<td>92</td>
</tr>
</tbody>
</table>

Note. Only those meta-analyses that provided a breakout for this construct were included.

timate derived from published studies and that derived from unpublished studies within the same set of meta-analyses. Published studies yielded mean effect sizes that averaged 0.14 SDs larger than unpublished studies. It is evident, therefore, that the treatment effects reported in published studies are indeed generally biased upward, relative to those in unpublished studies.

It is noteworthy, however, that the mean effect size estimates for both published and unpublished studies fall in the positive range; published studies are just more pos-
Examples of publication bias

• Medical journals from China almost never publish negative results (e.g. Pan et al. 2005)

• Only 5% of articles in journals focusing on Alternative and Complementary Medicine present negative results (Schmidt et al. 2001)

• Studies originating from Europe have more positive results than studies from the US (Sood et al. 2007)
CONCLUSION:
The efficacy of psychological interventions for depression has been overestimated in the published literature. Just as it has been for pharmacotherapy. Both are efficacious but not to the extent that the published literature would suggest.
Preregistering

• ClinicalTrials.gov

• Aims:
  – Increase transparency
  – Reduce fishing expedition bias
  – Presentation of post-hoc hypotheses as a priory
  – Enable assessment of publication bias
Publication bias assessment

Researcher bias against submitting negative results
Publisher bias against publishing negative results

Preregistering

- **PROSPERO**
  - Centre for Reviews and Dissemination
Examples

The Concept of a Systematic Review

- Studies
- Systematic review process
- Systematic Review
Internet-delivered CBT for insomnia

- Annual prevalence of insomnia: 10-20%
- 6% with a chronic trajectory
- Pharmacological treatment is non-curative and long-term use is associated with dependence, tolerance, side-effects, and increased mortality
- Cognitive Behavioral Therapy for Insomnia (CBT-I) is recommended as first choice – based on evidence from systematic reviews and meta-analyses
- Limited availability of CBT-I (trained therapists, geography, financial reasons)
- One possibility is Internet-delivered CBT-I (eCBT-I)
- *Is eCBT-I effective and are effects comparable to face-to-face delivered CBT-I?*

Vidensråd for forebyggelse (2015) Søvn og sundhed; American Academy of Sleep Medicine
Internet-delivered CBT-I

Statistically significant effects found for primary outcomes:

- Insomnia severity:
  Hedges’s $g = 1.09$, $p < 0.001$

- Sleep efficiency:
  Hedges’s $g = 0.58$, $p < 0.001$
Internet-delivered CBT-I

Comparing with face-to-face delivery

No statistically significant differences between internet-delivered and face-to-face-delivered CBT-I. Need for *non-inferiority trials* directly comparing eCBT-I and FtF.
Psychological intervention for distress in informal cancer caregivers

• Informal cancer caregivers (ICCs) report increased levels of psychological and physical morbidity and higher mortality

• Psychological interventions such as Cognitive Behavioral Therapies (CBTs) have been shown efficacious for distress (anxiety and depression)

• Aim: to evaluate the available evidence for the efficacy of CBTs for distress and physical symptoms among ICCs
Psychological intervention for distress in informal cancer caregivers

O’Toole, Zachariae, Penna, Mennin, Applebaum, 2016
Criticisms of meta-analysis
Criticisms of meta-analysis

"Exercise in mega-silliness" (H. Eysenck, 1978) ”Statistical alchemy” (Feinstein, 1995)

• Mixes apples and oranges
• Garbage in – garbage out
• File-drawer problem
• Reductionism (one number) $d = 2.7$ (95%CI: $1.7 - 3.7$)
## Systematic reviews

<table>
<thead>
<tr>
<th>Study Type</th>
<th>Transparency</th>
<th>Risk of bias</th>
<th>Effect estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-systematic review</td>
<td>Low</td>
<td>High</td>
<td>None</td>
</tr>
<tr>
<td>Narrative systematic review</td>
<td>Medium-high</td>
<td>Medium</td>
<td>Qualitative “Vote counting”</td>
</tr>
<tr>
<td>Quantitative systematic review (meta-analysis)</td>
<td>High</td>
<td>Low</td>
<td>Magnitude Direction Precision Sub-group comparisons</td>
</tr>
</tbody>
</table>
Lessons learned
“One must seek the truth where it is, not where one would like it to be”

Abbé de Faria, 1746-1819
Lessons learned

• When evaluating evidence
  – One study is not enough (replicability)
  – Avoid cherry picking – focus on the combined evidence of all available evidence
  – Less emphasis on p-values of individual studies – more emphasis on magnitude (effect size)
  – More emphasis on practical significance (e.g., MCID)
    • A highly statistically significant effect could be of a irrelevant magnitude
    • A non-statistically significant effect could potentially be clinically relevant
Lessons learned

• Consider
  – Statistical power and risk of Type-2 error
  – The precision of the effect (the confidence interval)
  – Homogeneity of the existing evidence
  – Study quality – potential bias and threats to validity
  – Publication bias – the tendency to underreport null-findings
  – Cost-effectiveness – relative to treatment as usual

• Establishing evidence
  – Is a complex cumulative process
Future tasks

• The future will bring public and policy-based demands for evaluation of clinical efficacy and cost-effectiveness

• Psychologists are advised to:
  – Work to establish a proactive, evidence-based professional culture
  – Not to take the effectiveness of psychological approaches as self-evident but to focus on the best available evidence
  – Accept when psychological approaches are not effective, be transparent about it, and work to improve the situation
  – Promote research-based practice and practice-relevant research, establish collaboration between researchers and clinicians, and conduct research-based evaluation in collaborative networks
Influence of Psychological Stress on Upper Respiratory Infection—A Meta-Analysis

Psycho-Oncology
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/pon.4144

Review
Cognitive behavioral therapies for informal caregivers of patients with cancer and cancer survivors: a systematic review and meta-analysis

Mia S. O'Toole1,*, Robert Zachariae2, Megan E. Renna3, Douglas S. Mennin4 and Allison Applebaum5
1Unit for Psychooncology and Health Psychology, Institute of Psychology and Behavioral Sciences, School of Business and Social Sciences, Aarhus University, Aarhus, Denmark
2Unit for Psychooncology and Health Psychology, Aarhus University and Aarhus University Hospital, Aarhus, Denmark
3Department of Psychology, CUNY Hunter College and the Graduate Center, New York, USA
4Department of Psychology, CUNY Hunter College and the Graduate Center, New York, USA
5Department of Psychology and Behavioral Sciences, Memorial Sloan Kettering Cancer Center, New York, USA

Abstract

Objective: Informal caregivers (ICs) of patients with cancer and cancer survivors report a number of psychological and physical complaints because of the burden associated with providing care. Given the documented effect of Cognitive Behavioral Therapy (CBT) on ICs' common psychological complaints, such as anxiety and depression, the objective was to conduct a meta-analysis on the effect of CBTs for adult ICs.

Methods: A literature search was conducted in order to identify all intervention studies on adult ICs that employed at least one therapeutic component defined as a CBT component.

Results: Literature searches revealed 36 unique records with sufficient data. These studies were subjected to meta-analyses using random effects models. A small, statistically significant effect of CBTs (Hedge's g = 0.08, p = 0.014) was revealed, which disappeared when randomized controlled trials were evaluated alone (g = 0.04, p = 0.200). A number of variables were explored as moderators. Only the percentage of female participants was positively associated with the effect size.

Conclusions: Based on the negligible effect of CBTs across outcomes, future studies should consider moving beyond traditional CBT methods as these do not appear efficacious. It is suggested that future interventions orient towards advances in the basic affective sciences and derived therapies in order to better understand and treat the emotional struggles experienced by ICs.

© 2016 The Authors. Psycho-Oncology published by John Wiley & Sons Ltd.