Assessing the Quality and Applicability of Systematic Reviews (AQASR)

Marcel Dijkers, PhD, FACRM
Icahn School of Medicine at Mount Sinai

Session 7 – April 2, 2014

An online workshop sponsored by SEDL's Center on Knowledge Translation for Disability and Rehabilitation Research (KTDRR)

Funded by NIDRR, US Department of Education, PR# H133A120012

Objectives:

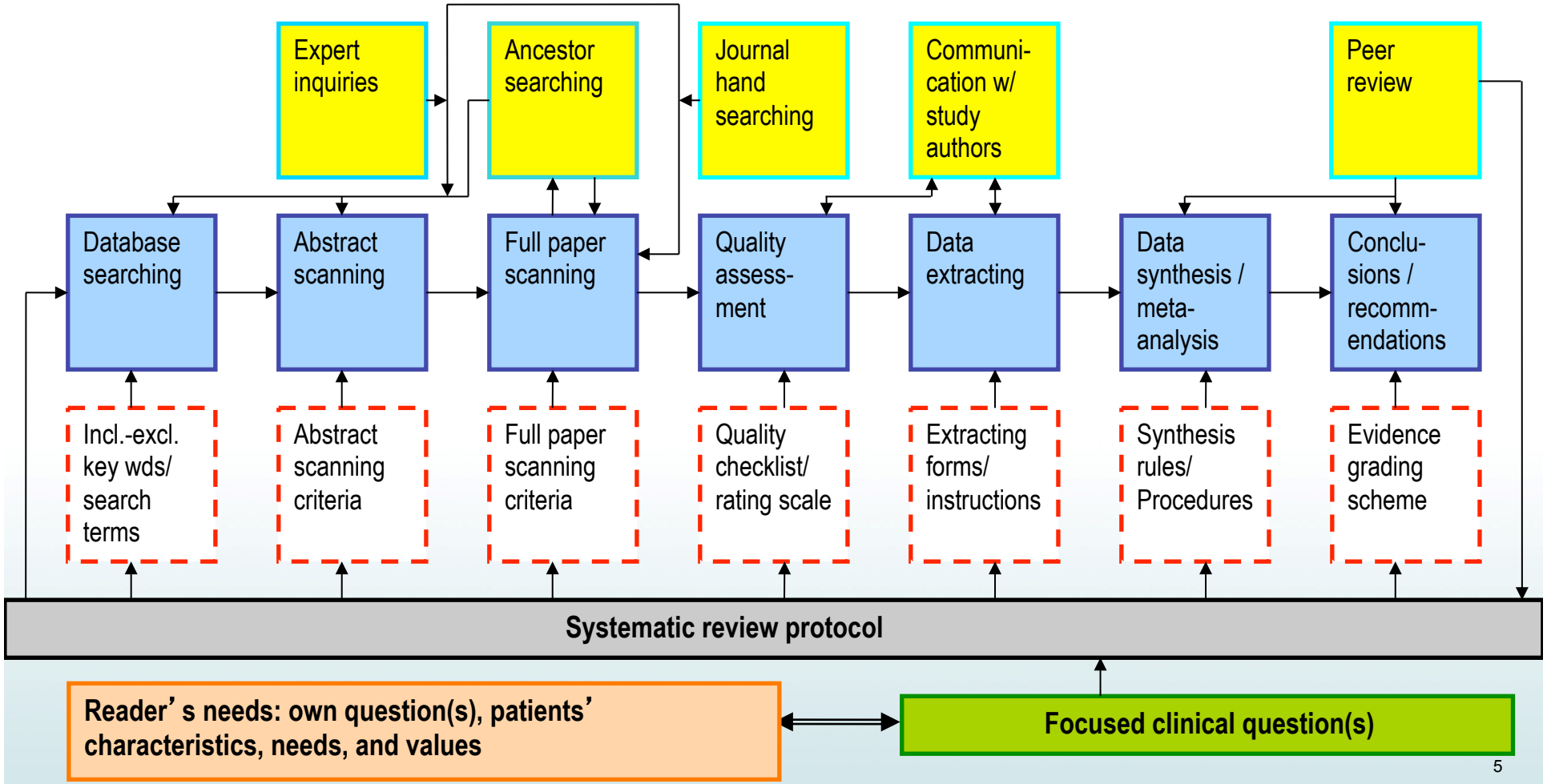
- Delineate steps and issues in the development of systematic reviews
- Introduce *Assessing the Quality and Applicability of Systematic Reviews (AQASR)* *(© SEDL/NCDDR 2011)*
- Describe how AQASR can be used in evaluating whether a particular systematic review can be trusted to provide an unbiased, reliable answer to one's (clinical, research, policy) question

Objectives:

- Review the various sections of AQASR and the items in each section
- Apply the instrument to several systematic reviews to increase familiarity with its elements and application

Questions?

The steps in a systematic review: schematic overview of systematic review production and the link of the results to the reader's interests



AQASR has questions on the steps all systematic reviews have in common:

- The focused clinical question (6)
- Systematic review protocol (5)
- Literature searches (16)
- Scanning of abstracts and full papers (8)
- Assessment of the quality of the primary studies (6)
- Extracting data (4)
- Synthesizing the data qualitatively (7)
- Drawing conclusions, making recommendations (7)

And (used by some)

- Synthesizing the data quantitatively (meta-analysis) (7)

In addition, AQASR has questions relevant to the topic of the systematic review:

- Intervention/prevention (13)
- Diagnostic procedure (8)
- Measurement instrument (10)
- Prognosis (6)
- Economic evaluation (7)

Questions?

Measurement instruments

MI1. Does the review describe the measure(s) reviewed, including content, unidimensionality vs. multidimensionality, number and nature of items, type of administration, equipment needed (if any), etc.?

- **Look for (1):** information in text or tables on basic characteristics of the measure(s), including:
  - developers and years of (re-)development
  - construct measured
  - subscales, if any, and number of items in each and overall
  - mode(s) of administration
  - potential for use of proxies
  - (original and later) target populations
  - (original and later) purpose
  - availability and source of norms

MI1. Does the review describe the measure(s) reviewed, including content, unidimensionality vs. multidimensionality, number and nature of items, type of administration, equipment needed (if any), etc.?

- **Look for (2):**
  - (A summary of) the definition of the construct(s) by the systematic reviewers, and by the authors of the primary studies or the scale's developers
  - A listing (in a table or appendix) of all or a sample of items of each of the instruments included in the review

MI1. Does the review describe the measure(s) reviewed, including content, unidimensionality vs. multidimensionality, number and nature of items, type of administration, equipment needed (if any), etc.?

- **Rationale:**
  - Systematic reviews of measurement instruments are written to assist clinicians and researchers in selecting instruments they can use in their work. The information on the measures reviewed is basic to understanding an instrument's characteristics and making a selection of one that is suitable for a particular application.

MI1. Does the review describe the measure(s) reviewed, including content, unidimensionality vs. multidimensionality, number and nature of items, type of administration, equipment needed (if any), etc.?

- **Dobson et al.: Measure class targeted**
  - physical functioning of people with hip/knee osteoarthritis (ICF Activities)
  - construct discussed by systematic review authors; aimed-for construct discussion by authors not addressed
  - performance based measures, quantification w/o expensive instruments
  - single-item (unidimensional) or multiple-item (unidimensional [reflective] or multidimensional [formative])

MI1. Does the review describe the measure(s) reviewed, including content, unidimensionality vs. multidimensionality, number and nature of items, type of administration, equipment needed (if any), etc.?

- **Dobson et al.: Specific measures**
  - items included noted in tables
  - mode of administration: performance (timing, counting, etc.) self-explanatory
  - proxies: N/A
  - norms: not noted

MI2. Does the review mention/discuss alternatives, especially older or better studied measures (possibly "gold standards"). Does the review address the role of the measure(s) of interest in the process of making decisions on clients/patients/subjects?

- **Look for:**
  - Information in text or tables on alternative measures for the same/closely related constructs, and their role in the systematic review (omitted, used as reference (validator) in some studies, etc.)

MI2. Does the review mention/discuss alternatives, especially older or better studied measures (possibly "gold standards"). Does the review address the role of the measure(s) of interest in the process of making decisions on clients/patients/subjects?

- **Rationale:**
  - Instruments that have a common term in their name (e.g., "quality of life") may differ widely in the construct operationalized, certainly in the definition and operationalization of a common construct. This affects comparability in terms of items included in the scales and in all psychometric qualities being considered. Instruments that are multidimensional in design or in actual functioning may need to be treated as two instruments.

MI2. Does the review mention/discuss alternatives, especially older or better studied measures (possibly "gold standards"). Does the review address the role of the measure(s) of interest in the process of making decisions on clients/patients/subjects?

- **Dobson et al.**
  - Alternatives: not discussed (self-report measures?)
  - "Gold standards": not discussed (except as applicable in validity assessment: comparison with PROs)
  - Role in clinical or research process: not discussed

MI3. Do the authors address the nature of the population sample(s) included in the primary studies, and the circumstances (testing conditions, etc.) in which psychometric information was collected?

- **Look for:**
  - Summary data on sample characteristics of all primary studies
  - Information on homogeneity and heterogeneity of these samples (within and between primary studies)
  - Information about the (dis)similarity of the sample(s) studied and the population the measure(s) in question are intended for or are commonly used for

MI3. Do the authors address the nature of the population sample(s) included in the primary studies, and the circumstances (testing conditions, etc.) in which psychometric information was collected?

- **Rationale:**
  - Psychometric characteristics, especially reliability and validity, are strongly affected by the nature and homogeneity of the sample. If the sample is atypical in terms of the population(s) from which it was drawn, a high reliability score may mean little, and similarly a low validity score may not be worrisome.

MI3. Do the authors address the nature of the population sample(s) included in the primary studies, and the circumstances (testing conditions, etc.) in which psychometric information was collected?

- **Dobson et al.:**
  - Osteoarthritis (knee/hip) was selection criteria; OA stage, age (mean and SD) reported when available; no other information on samples, data collection given
  - Homogeneity/heterogeneity of samples not addressed
  - Mismatch targeted population vis-à-vis studied samples: not discussed (no need?)

MI4. Do the authors assess the quality of the primary studies, including their size, completeness of data, and handling of missing data?

- **Look for (1):**
  - scoring of primary studies on methodological quality
  - justification of scale applied and its use
  - reference to a measure/measurement study rating system, such as COSMIN

MI4. Do the authors assess the quality of the primary studies, including their size, completeness of data, and handling of missing data?

- **Look for (2):**
  - a report of sample sizes
  - an evaluation of the representativeness of all samples of their purported population
  - a description of the research question(s) and hypotheses, if any, of the primary studies
  - data on the percentages of cases with a valid score for individual items
  - information on methods for handling missing information used by the primary studies

MI4. Do the authors assess the quality of the primary studies, including their size, completeness of data, and handling of missing data?

- **Look for (3):**
  - information on selective loss to follow-up, in longitudinal primary studies designed to measure sensitivity
  - an evaluation of the appropriateness of the statistical methods used in the primary studies
  - an evaluation of possible weaknesses or biases in the psychometric data reported that are due to other flaws in the design, implementation, analysis or reporting of the primary studies

MI4. Do the authors assess the quality of the primary studies, including their size, completeness of data, and handling of missing data?

- **Rationale:**
  - The reports of metric properties of the measure(s) included in a systematic review depend crucially on the quality of the primary studies. A reliable and useful systematic review should evaluate the primary studies that produced the estimates of validity, reliability and other psychometric characteristics the review synthesizes.

MI4. Do the authors assess the quality of the primary studies, including their size, completeness of data, and handling of missing data?

- **Dobson et al. (1):**
  - Study quality assessed using COSMIN tool
  - Sample size in table III, explicitly used in synthesis of findings of separate studies
  - Study representativeness not assessed (OA was an inclusion criterion)
  - Research questions/hypotheses of primary studies not addressed (purpose of assessment of at least one psychometric quality was inclusion criterion)
  - Data completeness of primary studies not addressed

MI4. Do the authors assess the quality of the primary studies, including their size, completeness of data, and handling of missing data?

- **Dobson et al. (2):**
  - Handling missing of information by primary studies not addressed (COSMIN?)
  - Loss to follow-up in responsiveness studies not addressed (COSMIN?)
  - Statistical analysis adequacy of primary studies not addressed (COSMIN?)
  - Other (sources of) weaknesses of primary studies not addressed (COSMIN?)

MI5. Does the review address the reliability/ reproducibility of the measure(s) included? If so, do the authors specify standards for what they consider minimally adequate reliability/ reproducibility? Was the application of these standards reproducible?

- **Look for:**
  - standards for adequacy listed in the text or the tables
  - a mention that no evidence regarding a particular reliability characteristic was available in the primary studies
  - evidence tables summarizing relevant reliability parameters from the primary studies

MI5. Does the review address the reliability/ reproducibility of the measure(s) included? If so, do the authors specify standards for what they consider minimally adequate reliability/ reproducibility? Was the application of these standards reproducible?

- **Rationale:**
  - A number of parameters for evaluating reliability are in existence, developed in various frameworks (for instance, internal consistency, inter-rater or intra-rater reliability in classical test theory; item separation reliability in Rasch analysis). Sometimes, standards for adequacy are set by the systematic review authors, based on suggestions in methodology textbooks (e.g., minimal adequate test-retest reliability is 0.70 for group applications, 0.90 for individual applications).

MI5. Does the review address the reliability/ reproducibility of the measure(s) included? If so, do the authors specify standards for what they consider minimally adequate reliability/ reproducibility? Was the application of these standards reproducible?

- **Dobson et al.:**
  - Internal consistency (if applicable), reliability and measurement error addressed (table III)
  - Standards provided (table I)
  - Straightforward application of the standards, presumably applicable

MI6. Does the review address the validity of the measure(s) included? If so, do the authors specify standards for what they consider minimally adequate convergent/ divergent and other types of validity? Was the application of these standards reproducible?

- **Look for:**
  - evidence tables summarizing relevant validity parameters (including correlations with a "gold standard") from the primary studies
  - standards for adequacy listed in the text or the tables
  - a mention that no evidence regarding a particular validity characteristic was available in the primary studies

MI6. Does the review address the validity of the measure(s) included? If so, do the authors specify standards for what they consider minimally adequate convergent/ divergent and other types of validity? Was the application of these standards reproducible?

- **Rationale:**
  - A number of parameters exist for evaluating validity of a scale, developed in various frameworks (for instance, construct, divergent and convergent validity in classical test theory; model fit statistics in Rasch analysis, information function in Item Response Theory). Sometimes, standards for adequacy are set by the systematic review authors. However, given the dependence of the parameters reported in the primary studies on the nature of the sample and the quality of other variables measured (e.g. the "gold standard" in construct validity), and the dependence of a judgment of "adequate" on one's conceptualization of the theory that links the construct of interest to other related and unrelated constructs, fixed standards are hard to defend. Certainly, the reproducibility of any judgments may be poor.

MI6. Does the review address the validity of the measure(s) included? If so, do the authors specify standards for what they consider minimally adequate convergent/ divergent and other types of validity? Was the application of these standards reproducible?

- **Dobson et al.:**
  - Content, structural, construct, cross-cultural and criterion validity (as available) addressed
  - Standards for validity given (table I); some of questionable relevance (e.g. “target population considers questionnaire to be complete“)
  - Judgment of satisfaction of most standards inherently somewhat subjective

MI7. Does the review address sensitivity of the measure(s) included? If so, do the authors specify standards for what they consider minimally adequate sensitivity?

- **Look for:**
  - evidence tables summarizing relevant sensitivity parameters from the primary studies
  - information on ceiling and floor effects, for all samples or for samples/ subgroups with the least/ most impairment
  - standards for adequacy of sensitivity listed in the text or the tables, including standards for the time elapsed between first and second assessments
  - a mention that no evidence regarding a particular sensitivity characteristic was available in the primary studies

MI7. Does the review address sensitivity of the measure(s) included? If so, do the authors specify standards for what they consider minimally adequate sensitivity?

- **Rationale:**
  - Sensitivity is a required characteristic for all measurement instruments used to assess change, whether that change is due to the natural history of a disease or results from interventions by rehabilitation clinicians. There are a number of parameters to express sensitivity, including the minimal detectable change, minimal clinically important difference, and the standardized mean difference. As time elapsed is a major determinant of the amount of change that can have occurred, all reported parameter values need to be evaluated in the light of the time elapsed between initial and subsequent assessments.

MI7. Does the review address sensitivity of the measure(s) included? If so, do the authors specify standards for what they consider minimally adequate sensitivity?

- **Dobson et al.:**
  - Responsiveness (as available) addressed (table IV)
  - Standard for responsiveness given (table I); parts are of questionable relevance
  - Judgment of satisfaction of standard (‘correlation with instrument measuring the ***same construct***’) inherently somewhat subjective

MI8. Does the review address the burden (cost, time, required skill levels, training, etc.) of collecting the data, imposed on the patients/ research subjects and/or on the researchers/ clinicians using the instrument?

- **Look for:**
  - information in the text or evidence tables on the burden issues most relevant to each type of measurement instrument
  - exact and approximate standards the systematic reviewers may use for "burdensome"
  - a mention that no evidence regarding administration burden was available in the primary studies
  - a section on costs, time and other burden issues, weighting them against the metric qualities of the scales

MI8. Does the review address the burden (cost, time, required skill levels, training, etc.) of collecting the data, imposed on the patients/ research subjects and/or on the researchers/ clinicians using the instrument?

- **Rationale**:
  - High-quality measures may be prohibitively expensive because of the cost of purchase or administration. These costs may include time (of administration and scoring), training, and risks (to subject/ patient and administrator). Good systematic reviews address these issues, and in making recommendations weigh costs against the value of the information produced by the measure(s) reported to have adequate psychometric qualities.

MI8. Does the review address the burden (cost, time, required skill levels, training, etc.) of collecting the data, imposed on the patients/ research subjects and/or on the researchers/ clinicians using the instrument?

- **Dobson et al.:**
  - 'No expensive instrumentation' was inclusion criterion
  - Equipment for each instrument listed
  - Time for subject and administrator not listed (generally minimal)
  - Skill/training of administrator not addressed
  - No discussion of 'costs' against psychometric qualities of instruments (because 'costs' are about the same across all?)

MI9. Do the reviewers offer a total score expressing their judgment of the overall quality of the instrument(s) included in their review? If so, do they specify which features of the instrument(s) played a role in formulating this overall judgment, and how? Do they make a clear distinction between lack of information and the availability of information that particular qualities are poor?

- **Look for:**
  - school letter grades (A through F, and U for insufficient information) in text or evidence tables
  - movie/restaurant review-type ratings (zero through five stars) in text or evidence tables
  - an explanation of the grading/rating system, including the basis on which reliability, validity and other psychometric qualities were weighed

MI9. Do the reviewers offer a total score expressing their judgment of the overall quality of the instrument(s) included in their review? If so, do they specify which features of the instrument(s) played a role in formulating this overall judgment, and how? Do they make a clear distinction between lack of information and the availability of information that particular qualities are poor?

- **Rationale:**
  - To simplify life for the users of measurement instruments, some systematic reviewers use a global rating for each of the scales reviewed, using various schemes for creating and expressing this global judgment. The final result depends very much on the psychometric and other qualities the authors emphasize, and users may not necessarily agree with their priorities. Certainly, reviewers ought to make the basis for their judgments as explicit as is possible.

MI9. Do the reviewers offer a total score expressing their judgment of the overall quality of the instrument(s) included in their review? If so, do they specify which features of the instrument(s) played a role in formulating this overall judgment, and how? Do they make a clear distinction between lack of information and the availability of information that particular qualities are poor?

- **Dobson et al.:**
  - Best measure in each category specified
  - No weighting of qualities specified
  - Lack of information specified in tables, and in summary by category; measures with lacking information not recommended

MI10. Do the authors address special issues of the use of the measure(s) by or with people with disabilities?

- **Look for (1):**
  - explicit statements that measures were included/ excluded or evaluated taking the needs of people with sensory, cognitive and other impairments into account
  - information in the text or evidence tables as to alternative methods of administration and their equivalence with the standard method
  - discussion of content (phrasing of items and response categories) that may be inapplicable, confusing or insulting to people with a disability

MI10. Do the authors address special issues of the use of the measure(s) by or with people with disabilities?

- **Look for (2):**
  - mention of special concerns as to the applicability and validity of the measure(s) to specific categories of people with disabilities, and/or summaries of the findings of the primary studies relevant to these issues

MI10. Do the authors address special issues relating to the use of the measure(s) by or with people with disabilities?

- **Rationale:**
  - Standardized tests may not be applicable to persons with a disability, and any conclusions based on the data may be invalid. Sensory and cognitive impairments may make it difficult for some categories of individuals to complete measures in their standard format. While alternatives are feasible (Braille, use of a reader, etc.), these may affect the quality of the instrument or the interpretation of findings. Some phrasing in instruments developed for the population at large may be incomprehensible or insulting to some categories of people with disabilities. Authors should address these and related issues that affect the feasibility of the instruments they review, and the interpretation of the data these produce.

MI10. Do the authors address special issues relating to the use of the measure(s) by or with people with disabilities?

- **Dobson et al.:**
  - All measures applicable to/tested in samples of people with OA
  - Limitations due to other disabilities (blindness, etc.) not addressed
  - Alternative methods of administration: not applicable
  - Content non-applicability, insulting, etc.: not applicable

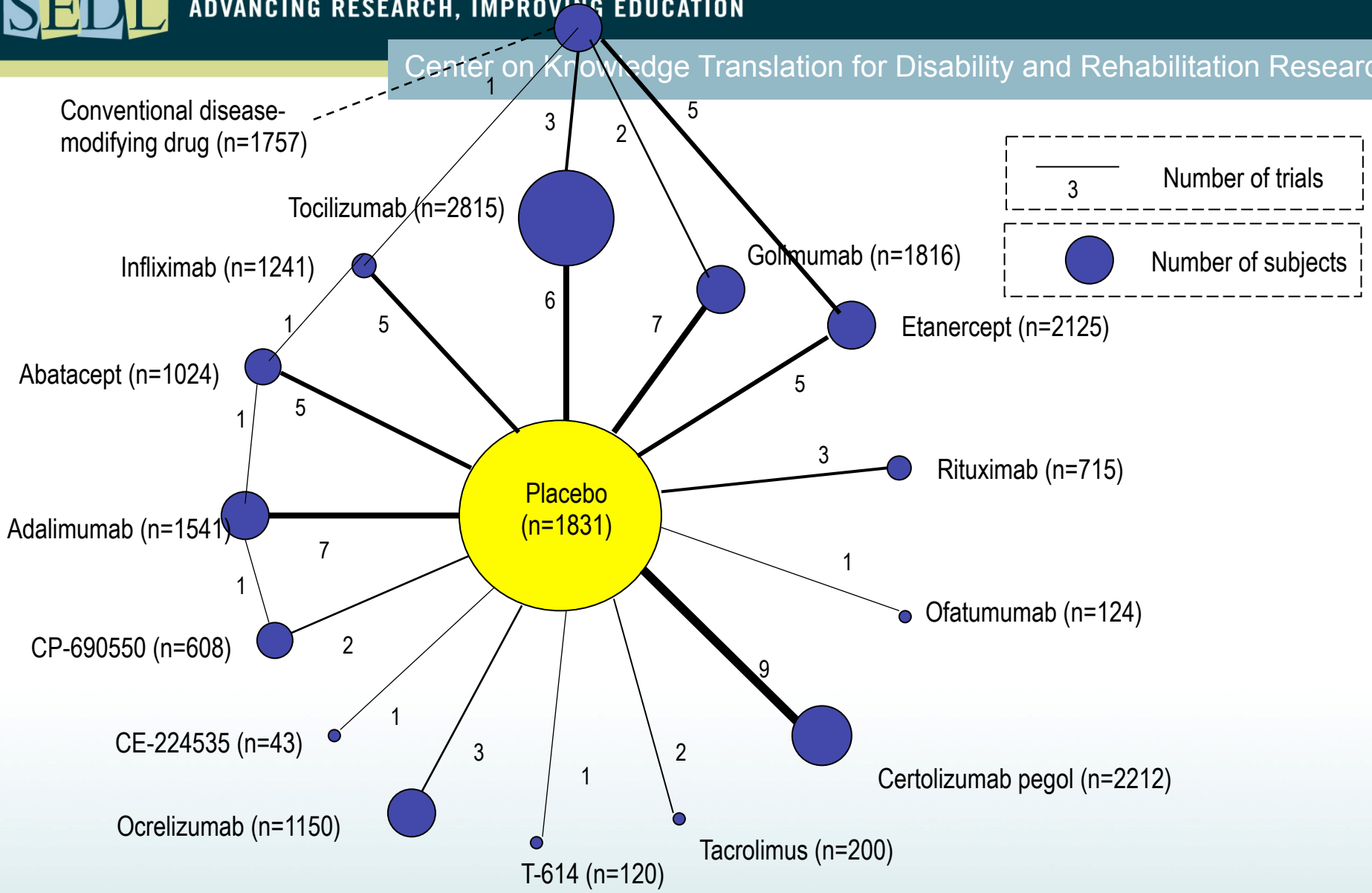
Questions?

AQASR: limitations

- Not applicable to:
  - Qualitative research synthesis
    - Meta-narrative review
    - Thematic analysis
    - Meta-ethnography
  - Mixed methods studies reviews
    - Mixed studies review
  - Scoping reviews

AQASR: limitations

- Not considering the specific issues of the latest developments in (quantitative) synthesis
  - Network meta-analysis (AKA indirect comparison)
  - Regression analysis with study as unit of analysis (meta-regression)
  - Individual patient (participant) meta-analysis



Network meta-analysis Redrawn after Estellat C, Ravaud P, Lack of Head-to-head Trials and Fair Control Arms: Randomized Controlled Trials of Biologic Treatment for Rheumatoid Arthritis. Arch Int Med 2012;172: 237-44

Regression analysis with study as unit of analysis (meta-regression)

Study	Mean age	% male	Baseline mean BDI	Mean # of sessions	% cured
Jones	45.8	72%	17.9	20.2	45%
Smith	29.4	81%	22.3	15.9	23%
Brown	33.4	56%	19.2	10.3	26.0
Onkky	58.0	55%	--	35.0	51%
Winky	--	76%	20.2	12.9	56%

Individual patient (participant) meta-analysis

- Cases harvested from the published literature (case studies, case series listing individuals)
- Data files obtained from other investigators
- Prospectively created sharing database

- Analysis for all three similar to that of (complex) primary studies

Oh et al.: Adjuvant radiotherapy delays recurrence following subtotal resection of spinal cord ependymomas (*Neuro-Oncology, 15*(2) 208-15, 2013)

- Searched literature for case studies/case series of spinal cord cancer treatment using surgery with/without radiotherapy: 68 articles describing 348 patients were found
- Extracted information on: age, sex, completeness of resection of tumor, disease recurrence or progression + time to it, adjuvant radiotherapy + total dosage, death + time to it, duration of follow-up
- Analyzed these data using Kaplan-Meier curves, Cox proportional hazards regression
- Conclusion: complete resection always better; with incomplete resection, radiotherapy helps significantly in slowing recurrence/progression, but not mortality

International Mission for Prognosis And Clinical Trial (IMPACT)

- Eight randomized placebo controlled trials and three observational studies conducted over the past 20 years
- 9205 patients with severe and moderate traumatic brain injuries
- Merged information on the pre-hospital, admission, and post-resuscitation assessments, acute management, and short- and long-term outcome
- ~20 reports on treatment and prognosis published

FITBIR: Federal Interagency Traumatic Brain Injury Research informatics system

- Next step after TBI Common Data Elements: database into which all investigators can "pour" their data
- Unique case identifier assigned based on name, birth date, place of birth
- Case identity invisible for database users, yet identifier allows identifying information from other studies that included the same case

- http://fitbir.nih.gov/

AQASR: alternatives: reporting guidelines

- QUOROM (*QU*ality *O*f *R*eporting *O*f *M*eta-analyses 1999 ~116 references)

re-created as:

- PRISMA (Preferred Reporting Items of Systematic reviews and Meta-Analyses) family (http://www.prisma-statement.org/):
  - PRISMA (original) (2009 ~370 references)
  - PRISMA for abstracts (of systematic reviews) (2013)
  - PRISMA Equity (2012)
- (reporting guidelines often used as standards so as to score quality of systematic reviews)

AQASR: alternatives: quality assessments

- AMSTAR (A MeaSurement Tool to Assess Reviews) family (http://amstar.ca/index.php): quality checking
  - AMSTAR (original) (2006; ~130 references)
  - R-AMSTAR (scores) (2010; ~8 references)
  - AMSTAR-NRS (non-randomized studies; under development)
- OQAQ: Overview Quality Assessment Questionnaire (1991; ~22 references)
- EVIDAAC systematic review scale (2008; 2 references)

AQASR: the future

- Hypertext version?
- Revised version with scoring system (cf. PRISMA)
- Updating to incorporate
  - New developments in systematic reviewing
    - Network meta-analysis
    - Regression meta-analysis
    - Individual participant meta-analysis
  - Aspects of conducting/reporting a systematic review not now included, that studies show to significantly impact outcomes

Systematic reviewing: the future

- Cochrane collaboration (Happy 20th birthday!)
- Campbell collaboration
- PROSPERO (http://www.crd.york.ac.uk/PROSPERO/)
- Ongoing methodological research
- Ongoing demand for systematic reviews
  - Needed by clinicians pursuing evidence-based practice
  - Prologue to a grant proposal for an RCT (e.g. PCORI)
  - Step in Clinical Practice Guidelines creation
  - Very good for journals' impact factor

WRAP-UP

Thank you for participating!

We invite you to:

- Provide your input on today's session
  - Workshop Evaluation form:

    http://survey.sedl.org/efm/wsb.dll/s/1g199
- Share your ideas for other webcasts
- Describe special needs you may have
- PLEASE CONTACT US:

joann.starks@sedl.org