Methodology Paper

Development of a quality appraisal tool for case series studies using a modified Delphi technique

March 2012

INSTITUTE OF HEALTH ECONOMICS
ALBERTA CANADA

INSTITUTE OF HEALTH ECONOMICS

The Institute of Health Economics (IHE) is an independent, not-for-profit organization that performs research in health economics and synthesizes evidence in health technology assessment to assist health policy making and best medical practices.

IHE BOARD OF DIRECTORS

Methodology Paper

Development of a quality appraisal tool for case series studies using a modified Delphi technique

Carmen Moga, MD, MSc
Bing Guo, MD, MSc
Don Schopflocher, PhD, MSc, BA (Honours)
Christa Harstall, BScMLS, MHSA

IHE Methodology Papers: Reports that provide information on health technology assessment topics with respect to methodological, policy, or administrative issues, but do not necessarily focus on published evidence. Production of this document has been made possible by a financial contribution from Alberta Health and Wellness. The views expressed herein do not necessary represent the official policy of Alberta Health and Wellness.

ACKNOWLEDGEMENTS

The authors would like to thank the following individuals for their participation in the Delphi process and provision of comments and suggestions for the proposed quality assessment tool.

- Ken Bond, IHE, Canada
- Alun Cameron, ASERNIP-S, Australia
- Paula Corabian, IHE, Canada
- Iñaki Imaz Iglesia, Spain
- Maria Ospina, IHE, Canada
- Ann Scott, IHE, Canada

The views expressed in the final report are those of the Institute of Health Economics.

The authors would like to thank Wendy McIndoo (IHE) for her support with checking references and formatting/editing the document and Ms. Patricia Chatterley and Ms. Liz Dennett, Information Specialists, Institute of Health Economics and University of Alberta, Edmonton, Alberta for searching the literature and information support.

Competing interest

The authors of this report declared no competing interests.

Corresponding Author

Please direct any inquiries about this report to Christa Harstall, charstall@ihe.ca

Funding

This report was supported by a financial contribution from Alberta Health and Wellness (AHW).

The views expressed herein do not necessarily represent the official policy of Alberta Health and Wellness.

Suggested Citation (ICJME or Vancouver Style)

Moga C, Guo B, Schopflocher D, Harstall C. *Development of a Quality Appraisal Tool for Case Series Studies Using a Modified Delphi Technique.* Edmonton AB: Institute of Health Economics. 2012.

Web Address

This publication is available for free download from the IHE website at: http://www.ihe.ca.

ABBREVIATIONS

AACPDM	American Academy for Cerebral Palsy and Development Medicine
ASERNIP-S	Australian Safety and Efficacy Register of New Interventional Procedures – Surgical
C	checklist
CDC	consensus development conference
CI	confidence interval
CONSORT	Consolidated Standards of Reporting Trials
CRD	Centre for Reviews and Dissemination
CS	case series study
E	excluded
G	generic
HTA	health technology assessment
ICC	Intra-class correlation
IHE	Institute of Health Economics
IQR	interquartile range
K value	Kappa value (interrater reliability)
MSAC	Medical Services Advisory Committee
N, n	number
NA	no response
NICE	National Institute for Health and Clinical Excellence
NRCS	nonrandomized comparative study
p	probability
R	reviewer
r	interclass correlation
RCT	randomized controlled trial
S	specific
SD	standard deviation
SE	standard error
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
W	Kendall's coefficient of concordance

TABLE OF CONTENTS

Acknowledgements i
Abbreviations ii
List of Figures and Tables v
Objectives and Scope 1
Background 1
Review of Critical Appraisal Tools for Case Series Studies 1
Development of a Quality Appraisal Checklist for Case Series Studies 5
General Characteristics of the Delphi Technique 5
Description of the Delphi Process 5
Selection of panelists (experts) 5
Delphi process: four rounds 5
Analyses of Responses 7
Results 7
Comparison With Another Published Checklist 8
Piloting the Newly Developed Quality Assessment Checklist in an HTA Project 10
Included Studies 10
Adaptation of the Draft Dictionary 10
Quality Rating Process 10
Results 11
Interrater reliability analysis 11
Variation in rating the included studies 12
Application of Quality Assessment Results 13
Discussion 14
Comprehensiveness of the Checklist 14
Feasibility and Usefulness of the Checklist 16
Interrater Reliability 17
Study Limitations 18
Future Improvements of the Checklist 18
Conclusion 18

Appendix A: Literature Search for Existing Quality Assessment Tools for Case Series Studies.....20
Appendix B: Excluded Publications 23
Appendix C: Results – Critical Appraisal Tools for Case Series Studies 27
Appendix D: List of Delphi Panelists 34
Appendix E: Results from Delphi Rounds 35
Appendix F: Results – Delphi First Round 37
Appendix G: Results – Delphi Second Round.....40
Appendix H: Results – Delphi Third Round.....44
Appendix I: Criteria and Draft Dictionary for the Quality Assessment Checklist.....46
Appendix J: Adaptation of the Draft Dictionary 49
Appendix K: Suggestions for the Checklist and Draft Dictionary 52
Appendix L: Preliminary Factor Analysis.....57
References 62

LIST OF FIGURES AND TABLES

Figure L.1: First dimension of study by criterion/item matrix (SigmaPlot® 11) 57

Figure L.2: Second and third factors of variation (SigmaPlot® 11) 58

Figure L.2.1 - Figure L.2.4: Second and third dimensions of variation in the study by criterion/item matrix (SigmaPlot® 11) 59

Table 1: Identified tools for the critical appraisal of case series studies 2

Table 2: Methods used to develop the checklists and validate the published tools 3

Table 3: Frequency of criteria included in the published checklists 4

Table 4: Comparison of two checklists 9

Table 5: Summary Kappa – three reviewers 13 studies, 18 criteria 11

Table 6: Summary Kappa – three reviewers versus their consensus scorings 11

Table 7: Means and standard deviations of the rating for each study across the three reviewers, 18 criteria 12

Table 8: Means and standard deviations of ratings for each criterion across three reviewers, 13 studies 13

Table 9: Comparison of commonly used criteria found in the published checklists for case series studies only and the 18-criteria checklist 14

Table A.1: Search strategy 20

Table B.1: Excluded publications 23

Table C.1: Methods used to develop the checklists and internal validation of the published tools 27

Table C.2: Criteria from checklists used for the appraisal of case series studies 29

Table E.1: Summary of results from Delphi rounds 35

Table F.1: Panelists' ranking (first round) 37

Table F.2: Panelists' responses (first round) 38

Table F.3: Suggestions made by panelists (first round) 39

Table G.1: Panelists' ranking (second round) 40

Table G.2: Panelists' responses (second round) 41

Table G.3: Results and decision for suggestions made by panelists in the first round 42

Table H.1: Quality appraisal criteria selected from the first two rounds 44

Table H.2: Quality appraisal criteria excluded from the first two rounds 45

Table J.1: Adaptation of the draft dictionary 49

Table K.1: Suggested modifications to the checklist and draft dictionary 52

OBJECTIVES AND SCOPE

This information paper outlines the process undertaken by a group of researchers at the Institute of Health Economics (IHE) in collaboration with researchers from two other health technology assessment (HTA) agencies in Australia and Spain to develop a checklist for quality appraisal of case series studies using a modified Delphi technique.

In addition to this work, a brief review of other published checklists was undertaken, and the results of a pilot test of the newly developed quality appraisal checklist are presented.

BACKGROUND

A case series is an observational study describing a series of individuals, usually all receiving the same intervention with no control group.[1] Because of the lack of a control group, a case series study occupies a low position in the hierarchy of evidence and is considered the weakest study design from which to obtain evidence on effectiveness. Case series studies may be affected by various types of biases related to selection, detection, performance, attrition, reporting, and publication. Thus the derived results are ranked as low quality.[2] Nevertheless, there are circumstances when case series studies are the only form of research evidence available and including them in systematic reviews and HTA reviews might be considered necessary.[3]

No universally accepted quality appraisal tool exists for assessing the methodological quality of case series studies.[4] Several reviews have been conducted to identify instruments that assess the quality of nonrandomized studies of health interventions, including case series studies. Saunders et al.[5] in 2003 found that in the reviewed instruments published up to March 1999, there was a great variation in terms of their scope, the number and nature of the criteria included, and the rigour of their development. Mallen et al.[4] in a 2006 publication concluded that quality assessment does not routinely occur in systematic reviews of observational studies and, where it does occur, there is no clear consensus on the method used. The same authors found that quality assessment was conducted in 22% of the systematic reviews published between 1999 and 2000, compared to 50% of reviews published between 2003 and 2004.[4] More specifically, in an HTA review published in 2005, Dalziel et al.[2] stated that there was no consensus on which case series studies to include in HTAs, how to use them, or how to assess their quality, despite the fact that such studies have been used in 30% of the HTAs produced by the National Institute for Health and Clinical Excellence (NICE).

Review of Critical Appraisal Tools for Case Series Studies

A literature search was conducted to identify studies published in English between January 1998 and June 2011 on the development or use of tools designed to assess the quality of case series studies (Appendix A, Table A.1). Study selection was conducted by one reviewer based on study abstracts and/or the full-text articles. The publications were included if they mentioned the development and/or use of a quality appraisal tool. Quality appraisal tools used for assessment of randomized controlled trials, nonrandomized comparative studies, or case series studies were also included if the authors explicitly mentioned that they used those tools to appraise case series studies. Publications in which authors graded the studies only on the basis of the study design but did not apply a quality appraisal tool were excluded. The list of excluded studies and the reasons for exclusion are presented in Appendix B, Table B.1. Data were extracted and synthesized qualitatively by one reviewer.

Thirty-six studies were found by the literature search (Table 1). Ten studies[6-15] included checklists used to appraise case series studies only, whereas 26 used checklists to appraise studies of various designs.[16-41] One third of the studies were HTA publications. The number of criteria included in the checklists varied widely, ranging from 3 to 30 for checklists used to appraise case series studies (median 8.5; IQR: 6; 13.75) and 4 to 61 criteria for checklists used to appraise various study designs (median 15.5; IQR: 11.25; 24) (Table 1).

Table 1: Identified tools for the critical appraisal of case series studies

Study[a]	No. criteria	Generic or specific[§]	Type of instrument	Study type assessed	
				CS	RCT or NRCS
1. Young et al.[14] 1999*	3	G	Checklist	✓	–
2. McCrory et al.[9] 2001*	5	S	Checklist	✓	–
3. CRD's Guidance[7] 2001*	6	G	Checklist	✓	–
4. Taylor et al.[12] 2005	6	G	Checklist	✓	–
5. National Collaborating Centre for Acute Care[11] 2003*	8	G	Checklist scale	✓	–
6. Chipchase et al.[15] 2009	9	G	Checklist	✓	–
7. Yang et al.[13] 2009	13	G[‖]	Checklist	✓	–
8. Cauchi et al.[†6] 2008	14	G	Checklist	✓	–
9. Huisstede et al.[8] 2008	15	G	Checklist	✓	–
10. Moga & Harstall[10] 2006*	30	G	Checklist scale	✓	–
11. Stead et al.[34] 2008	4	G	Checklist scale	✓	✓
12. AACPDM methodology[16] 2008* version	7	G	Checklist	✓	✓
13. Mortenson & Eng[29] 2003	7	G	Checklist	✓	✓
14. Overend et al.[31] 2001	8	G	Checklist	✓	✓
15. Bryant et al.[17] 2002*	9	G	Checklist	✓	✓
16. Wells et al.[41] 2000	9	G	Checklist	✓	✓
17. PEDro Scale[32] 2009	11	G	Checklist scale	✓	✓
18. Oremus et al.[30] 2008*	12	S	Checklist	✓	✓
19. Slim et al.[33] 2003	12	G	Checklist scale	✓	✓
20. Thomas et al.[36] 2006	13	S	Checklist scale	✓	✓
21. Smith T et al.[39] 2008	14	G	Checklist	✓	✓
22. Steward et al.[35] 1999*	14	S	Checklist	✓	✓
23. Deenadayalan et al.[40] 2010	15	G	Checklist	✓	✓
24. Haines et al.[23] 1999	16	S	Checklist	✓	✓
25. Merlin et al.[28] 2001*	16	G	Checklist	✓	✓
26. Reiman et al.[38] 2009	19	G	Checklist	✓	✓
27. Des Jarlais et al.[20] 2004	22	G	Checklist	✓	✓

28. Hayashi et al.[25] 2003	24	G	Checklist	✓	✓
29. MacDermid[27] 2003	24	G	Checklist scale	✓	✓
30. Nichol et al.[18] 1999[¶]	24	G	Checklist scale	✓	✓
31. de Kleuver et al.[19] 2003	25	G	Checklist	✓	✓
32. Helm et al.[37] 2009	26[#]	G	Checklist	✓	✓
33. Downs & Black[21] 2001	27	G	Checklist scale	✓	✓
34. Jongerius et al.[26] 2003	27	G	Checklist scale	✓	✓
35. Green et al.[22] 1999*	29	G	Checklist	✓	✓
36. Hardy et al.[24] 2001*	61	S	Checklist	✓	✓

[a] Studies are ordered by number of criteria included and study type assessed.

* HTA and other reports; [†] checklist was developed by the Review Body for Interventional Procedures, UK; [§] specific: included criteria tailored for a particular medical condition; [‖] checklist was developed for appraisal of studies on herbal medicine, but the criteria are written in a generic form; [¶] used a validated checklist published by Cho et al.[42] in 1994; [#] criteria are organized into nine domains

Abbreviations: AACPDM: American Academy for Cerebral Palsy and Development Medicine; CRD: Centre for Reviews and Dissemination; CS: case series study; G: generic; No: number; NRCS: nonrandomized comparative study; RCT: randomized controlled trial; S: specific; ✓: yes; −: no

Twenty-one checklists are adaptations or modifications from other sources,[6-8,10,15,17-19,21,23,25,29,30,34,36-40,42,43] while the authors did not describe the process for the development of the checklists in 13 publications[9,11,14,20,22,24,26-28,32,33,35,41] (Appendix C, Table C.1). Only two publications indicated that experts or authors of the publications developed the checklists.[13,31] Eleven publications[6,7,13,15,16,18,19,21,25,31,33] provided some details about the methods used to select the criteria and to develop the checklists. In six publications[10,13,18,21,25,33] the researchers measured the interrater reliability of the criteria included in the checklists, three[13,18,21] provided estimated times for completion of the appraisal using the checklist, and 15 provided instructions for scoring the checklist criteria.[8,10,13,14,16-18,21,22,27,32-34,36,41]

Details on the methods used to develop the checklists and on the internal validation of the published tools were reported in only three studies (Table 2). One of the checklists[13] was specifically developed for the evaluation of herbal medicine case series studies.

Table 2: Methods used to develop the checklists and validate the published tools

Study	Source of the tool	Method used to develop the checklist (criteria selection, process)	Interrater reliability reported	Instructions provided for scoring	Time for completion
Yang et al.[13] 2009	Developed by experts	A panel of experts generated initial criteria; modified Delphi technique – rated importance of criteria; pretested; refinement of the instrument	Cronbach's α between 0.80 and 0.85 ICC: 0.904 (high)	Yes	≤ 15 minutes per paper
Nichol et al.[18] 1999[†]	Modified from Spitzer et al.[44‡]	Eliminated criteria which did not address systematic bias and consistency; pretest instrument[‡]	W = 0.64 r = 0.89 (95% CI 0.73 to 0.93)[‡]	Yes[‡]	Approx. 30 minutes per article/ reviewer[‡]

Downs & Black[21] 1998	Developed from other checklists used for the assessment of RCTs[45-51]	Pilot test followed by revisions; measurement of internal consistency, test-retest reliability, interrater reliability, face and criterion validity	r = 0.75 (good)	Yes	Average 20 to 25 minutes; range 10 to 45 minutes per paper

* Grey shaded row indicates checklist developed for appraisal of case series studies only; † used a validated checklist published by Cho et al.[42] in 1994; ‡information abstracted from Cho et al. 1994.[42]

Abbreviations: ICC: Intra-class correlation; r: interclass correlation; W: Kendall's coefficient of concordance.

An inventory of all the criteria included in the 36 checklists is listed in Table C.2, Appendix C. The criteria are organized within six domains: study question, study population, intervention, outcome measurement, statistical analysis, and results.

Criteria which were included in at least two studies are provided in Table 3 and are ordered by frequency of inclusion. These criteria cover the six domains previously mentioned.

Table 3: Frequency of criteria included in the published checklists

Number of studies (N = 36) / Criterion*	Number of studies that used checklists only for CS (rank) N = 10	Number of studies that used checklists for CS and studies of other design (rank) N = 26
Clear definition of the primary and secondary outcomes	8 (1)	18 (1)
Description of the eligibility criteria (inclusion and exclusion criteria)	8 (1)	15 (3)
Relevant/accurate outcomes reported	6 (2)	8 (7)
Blind assessment of outcomes	5 (3)	18 (1)
Appropriate methods for recruitment of participants (adequate sample, relevant population)	5 (3)	12 (4)
Duration of follow-up reported and appropriate	5 (3)	10 (6)
Prospective study design	5 (3)	4 (11)
Description of the purpose, aim, or objectives of the study	4 (4)	10 (6)
Participants recruited consecutively	4 (4)	4 (11)
Report of loss to follow-up	4 (4)	17 (2)
Description of the intervention	4 (4)	12 (4)
Statistical tests appropriate/valid	3 (5)	11 (5)
Description of adverse events/ side effects	3 (5)	8 (7)
Participants entering the study at a similar point in their disease progression	3 (5)	–
Description of co-intervention(s) received	3 (5)	4 (11)
Description of baseline characteristics of participants such as age and gender	2 (6)	8 (7)
Objective methods to measure outcomes	2 (6)	2 (13)
Recruitment period clearly stated; same time	2 (6)	–
Case series included more than one centre (multicentre study)	2 (6)	–

* Criteria included in at least two studies which used checklists to appraise CS studies only

Abbreviations: CS: case series study; N: number

DEVELOPMENT OF A QUALITY APPRAISAL CHECKLIST FOR CASE SERIES STUDIES

General Characteristics of the Delphi Technique

The Delphi technique is a consensus development method used extensively in health care.[52,53] It was developed by the RAND Corporation in the 1950s, where it emerged as a method of eliciting and refining group judgments. The Delphi technique is a rapid and relatively efficient way to collect information from a group of knowledgeable people (panel) by taking into consideration the opinion of each member on the panel.[52]

The technique has three features: response anonymity (it allows a sharing of responsibility and releases responders' inhibitions), iteration (processes occur in rounds) and controlled feedback (showing the distribution of the group's response), and statistical aggregation of group responses (expressing judgment using summary measures of the full group response).[52] This method was chosen over other consensus techniques because of its ability to allow all group members equal participation and influence, even when separated geographically.[54] The results of a Delphi exercise are more readily accepted than are those obtained by consensus or by more direct forms of interaction.[52]

Description of the Delphi Process

A modified Delphi technique was employed to further refine a checklist for the appraisal of the quality of case series studies.

Selection of panelists (experts)

The panel consisted of seven HTA professionals self-selected from the following institutions/agencies:

- Institute of Health Economics (IHE), Canada: five panelists;
- Australian Safety and Efficacy Register of New Interventional Procedures – Surgical (ASERNIP-S), Research & Academic Surgery Division, Australia: one panelist;
- Agency for Health Technology Assessment, Institute of Health "Carlos III", Ministry of Science and Innovation, Spain: one panelist.

The panelists are a homogeneous group with an intimate knowledge of the various critical appraisal tools used in the field of HTA and the benefits and strengths of using case series studies as an evidence base. Experts involved in the conduction of primary research (i.e. case series, clinical trials) were not included in the panel. A list of the panelists is provided in Appendix D.

Delphi process: four rounds

The objectives of the four Delphi rounds are outlined in Box 1.

As with a modified Delphi technique, the panel did not participate in an initial round usually undertaken in a Delphi study to compile a list of criteria. Instead, the panel rank ordered a previously composed inventory.[10] Two researchers at the IHE developed the initial checklist of 30 criteria using criteria from five studies[2,11,21,55,56] identified through a limited search of literature.

During the first round, the panelists had the opportunity to suggest criteria that were not included in the initial checklist.

Box 1: Objectives of the four rounds

First round:
- To rank the importance of each criterion included in the initial checklist
- To suggest new criteria, if needed

Second round:
- To provide feedback on the results of the first round
- To re-rank the importance of the criteria which did not reach 70% agreement for inclusion or exclusion
- To indicate if the new criteria proposed during the first round should be included

Third round:
- To further refine the checklist and exclude any criterion considered less important or to re-include in the checklist any of the excluded criteria

Fourth round:
- To review the final checklist and draft dictionary for further improvement

A four-stage e-mail-based modified Delphi process conducted between November 2006 and April 2007 was used to cull the initial checklist of 30 criteria to a more "user friendly" checklist, as follows.

The appropriateness of each criterion was rated on a five-point Likert scale that ranged from 1 (very important) to 5 (not important at all), with one equivocal point 3, and two "grey" points 2 and 4 (Box 2).

Box 2: Ranking definitions

Rank 1: Very important. You are confident that the criterion should be included in the appraisal checklist.

Rank 2: Somewhat important and perhaps should be included.

Rank 3: Equivocal. You are not sure if the criterion should be included or excluded.

Rank 4: Not very important and perhaps should be excluded.

Rank 5: Not important at all. You are confident that the criterion should be excluded.

The panelists received an e-mail describing the Delphi process and the expectations regarding their participation. The questionnaires were sent electronically. Participants were given 2 to 3 weeks to respond to each questionnaire. Although participants were aware of the identities of other responders, they were blind to individual responses, ensuring anonymity throughout the process. At the beginning of the process, a random number from 1 to 7 was allocated to each panelist, and these numbers were maintained throughout the process. The responses were analyzed anonymously by a biostatistician, an expert in instrument development who was not a member of the panel and was blinded to the identities of panel members. Results and suggestions for new criteria made during the first round were summarized and returned to the panelists for further consideration in round 2. Panelists were not asked to comment on the reasons for including or excluding criteria from the list. Each panelist received a personalized summary of their own results and the results and distribution

of ratings assigned by the other panelists in the previous round. This procedure allowed each participant to see his/her own and the aggregate group's ratings.[57]

The responses from all Delphi rounds were compiled by two panelists who were blinded to the identities of other panelists, with assistance from an independent assistant, who collated and communicated all the feedback from and to the panelists to preserve confidentiality.

Analyses of Responses

A criterion was considered appropriate for inclusion or exclusion in the final quality appraisal checklist if at least five out of seven panelists judged that criterion very important (rank 1) or not important (rank 5), showing a 70% agreement among panelists. This cut-off was decided a priori. The same approach was repeated for rounds 2 and 3. Data from all Delphi rounds were analyzed quantitatively by an independent biostatistician, an expert in instrument development.

Two panelists developed a draft dictionary for the checklist which was shared with the other panelists in round 4. The original version of the checklist included three levels of responses for each criterion (yes, no, unable to determine). To simplify the rating process, dichotomous qualitative categories (yes and no) were used for responses. The response "unclear/unable to determine" was collapsed into the no response category although it was known that this approach might underestimate the reported characteristics.

Results

The results of the Delphi rounds are presented in Appendix E. A 100% response rate was reached in the first three rounds.

First round

Fourteen of the 30 criteria were judged very important (rank 1) by at least five panelists (70% agreement). Sixteen criteria which received ranking scores lower than 1 were excluded. The panel suggested to combine existing criteria that appeared similar, include three new criteria, and to increase the specificity and clarity of some criteria (Appendix F). The included and excluded criteria and the summary of suggestions and comments were sent back to the panelists for the second round.

Second round

Panelists had the opportunity to further review the 16 criteria not included during the first round. Five of those criteria were judged very important by 70% of the panelists and were added to the final checklist. Eleven remaining criteria were excluded (Appendix G).

All panelists agreed that the checklist should be used for intervention studies only (with a before-and-after comparison), and six out of seven participants considered that the checklist should include explanations for each criterion. None of the three new criteria suggested in the first round met 70% agreement to be included in the final checklist. The 19-criteria checklist and 11 excluded criteria were sent back to the panel for review.

Third round

One of the 19 included criteria initially considered important was voted for exclusion by five panelists (70% agreement). Some criteria were slightly reworded to improve their consistency and

clarity (Appendix H). The final 18-criteria checklist along with a draft dictionary explaining each criterion was sent back to the panelists for their review and comments (Appendix I).

Fourth round

No further communication or comments were received from the panelists, and the checklist and draft dictionary were accepted in the submitted form.

Comparison with Another Published Checklist

The literature search identified one checklist published by Yang et al.[13] in 2009 that was developed through a comprehensive process, including a Delphi technique. The checklist was specifically constructed for the evaluation of case series studies in the field of herbal medicine. The authors described a two-stage process of developing the checklist: the checklist was initially generated by judges (experts in health care practice, research, instrument development, or a combination of these), and then was improved and validated by judges and raters (herbal medicine researchers, herbal medicine practitioners, and other academics in other health care disciplines) by using the draft instrument to assess the quality of 47 case series studies. The initial version of the instrument included 68 criteria identified by experts during a modified Delphi technique. A content validity assessment (consensus of the judges) reduced the number of criteria to 24. These criteria were used to assess 12 case series studies in the pretest stage. The sequence of the items was then reorganized, and each item was transformed from a phrase to a narrative sentence based on the raters' recommendations from the pretest. The final test included the assessment of reliability and construct validity of the 24-criteria checklist on a sample of 35 case series studies on Chinese herbal medicine, resulting in a further reduction of the number of criteria to 13. Table 4 shows a comparison of the criteria included in the 18-criteria checklist developed in this project with Yang et al.'s[13] checklist of 13 criteria.

Although the 13-criteria checklist includes fewer criteria, some of the criteria cover more than one aspect. For example, criterion 5 covers several aspects of the treatment protocol, such as intervention and outcome measures, and criterion 9 focuses on definitions of therapeutic effects and side-effects. Six criteria are common to both checklists: clear aim of the study, explicit/clear inclusion and exclusion criteria, clear description of the intervention, objective assessment, appropriate data/statistical analysis, and reported adverse events. Two additional descriptive criteria from the 18-criteria checklist that focus on recruitment of participants (criteria 5 and 6) were captured in one criterion in the 13-criteria checklist (criterion 7). Terms used in both checklists such as *appropriate*, *adequate*, or *relevant and complete* need clarification, however. The checklist published by Yang et al. [13] did not include a dictionary. The 13-criteria checklist includes six new broad criteria (i.e. criteria 2, 3, 4, 8, 11, and 12) which resemble some of the criteria included in the 18-criteria checklist. Overall, no important criterion was missing in our 18-criteria checklist when compared to the 13-criteria checklist.

Table 4: Comparison of two checklists

18-criteria checklist by the Delphi panel*	13-criteria checklist by Yang et al.[13]*
Study objective	
1. Is the hypothesis/aim/objective of the study stated clearly in the abstract, introduction, or methods section?	1. The rationale/aim of the study is clear.
Study population	
2. Are the characteristics of the participants included in the study described?	–
3. Were the cases collected in more than one centre?	–
4. Are the eligibility criteria (inclusion and exclusion criteria) for entry into the study explicit and appropriate?	6. Inclusion/exclusion criteria (age range, disease/symptom duration, selection endpoints, diagnosis) are clear.
5. Were participants recruited consecutively?	7. The methods of patient recruitment are appropriate.
6. Did participants enter the study at a similar point in the disease?	
Intervention and co-intervention	
7. Was the intervention clearly described in the study?	5. The treatment protocol (intervention and its duration, outcome measures: qualitative or quantitative, long-term vs. short-term, endpoints) is adequately described.
8. Were additional interventions (co-interventions) clearly reported in the study?	–
Outcome measure	
9. Are the outcome measures clearly defined in the introduction or methods section?	–
10. Were relevant outcomes appropriately measured with objective and/or subjective methods?	10. Subject assessment was independent and objective.
11. Were outcomes measured before and after intervention?	–
Statistical analysis	
12. Were the statistical tests used to assess the relevant outcomes appropriate?	13. Data analysis is appropriate for the design of the study.
Results and conclusions	
13. Was the length of follow-up reported?	–
14. Was the loss to follow-up reported?	–
15. Does the study provide estimates of the random variability in the data analysis of relevant outcomes?	–
16. Are adverse events reported?	9. Therapeutic effects and side-effects are defined.
17. Are the conclusions of the study supported by results?	–
Competing interests and sources of support	
18. Are both competing interests and sources of support for the study reported?	–

–	2. Description of the disease/condition being treated is adequate.
–	3. The study design is appropriate for the aim of study.
–	4. The rationale for the treatment protocol is clear.
–	8. Details of methods/ procedures are adequate to allow the study to be repeated.
–	11. The results for all outcome measures have been clearly reported.
–	12. The data collected are relevant and complete.

* The assigned number for each criterion is the same as in the original publication.

PILOTING THE NEWLY DEVELOPED QUALITY ASSESSMENT CHECKLIST IN AN HTA PROJECT

The 18-criteria checklist developed through the Delphi process was piloted in an HTA project[58] that assessed the clinical research evidence on the safety and efficacy/effectiveness of islet transplantation in patients with type 1 diabetes.

Included Studies

Research evidence on the safety and efficacy/effectiveness of islet transplantation comes almost exclusively from case series studies with a before-and-after comparison.

Although case series studies are considered the weakest study design to examine effects of a treatment because of the lack of a control group, in this review they were the only source of research evidence available. Furthermore, even without a parallel comparison, the association between the treatment (islet transplantation) and some outcomes such as insulin independence can still be established by a before-and-after comparison because insulin independence can be achieved only by the treatment (by definition, insulin-dependent diabetes). In other words, the impact of placebo effects would be minimal in this case.

Adaptation of the Draft Dictionary

Prior to conducting the quality assessment, one reviewer customized the dictionary for seven criteria (criteria 2, 4, 6, 7, 8, 9, and 12) by incorporating some important clinical aspects relevant to the patients and the treatment of interest (Appendix J). Two reviewers who were familiar with the research related to the treatment for diabetes discussed and agreed on the adaptation. The majority of the criteria with the adapted dictionary were easy to apply; the dictionary for criterion 6, however, underwent several modifications (Appendix J).

Quality Rating Process

Two reviewers (R1, R2) independently appraised the methodological quality of the 13 case series studies[59-71] using the 18-criteria checklist with the adapted dictionary. The ratings by the two reviewers were then compared; disagreements were resolved by consensus.

To further examine the interrater reliability of the checklist, a third reviewer (R3) who was not involved in the discussion and adaptation of the dictionary also appraised the methodological quality

of the same studies using the modified dictionary. The rating results of the three reviewers were compared to each other as well as to the consensus results. Disagreements were discussed in detail among the three reviewers.

Results

Interrater reliability analysis

An interrater reliability analysis using the Kappa statistic (SPSS 15.0) was conducted to determine consistency of scoring among the three reviewers (R1, R2, R3). The results showed different levels of agreement among them. A better agreement was achieved between R1 and R2 than between R3 and either of the other two reviewers (Table 5).

The ratings allocated by each of the three reviewers were further compared with the consensus ratings obtained after the discussions and resolution of disagreements. Overall the consensus ratings obtained from all three reviewers were in substantial agreement with the ratings made by R1 and R2 individually (Kappa = 0.806) and were in moderate agreement with R3 (Kappa = 0.552) (Table 6). These results speak to the importance of involving multiple reviewers in the appraisal process and also to the important role of preliminary discussions, clarifications, and potential adaptation of the appraisal tool and its dictionary before beginning the appraisal, which is vital for increasing the clarity and ultimately the level of agreement in scoring the criteria in the checklist. Ad hoc involvement in the appraisal process of additional reviewers who do not always have an intimate link with the subject reviewed (such as reviewer R3) is common in practice, often triggered by sparse resources and the requirement for a well-conducted review to involve at least two independent reviewers in the quality assessment of the research. Including a third reviewer provided a glimpse into how the new checklist might work when it is used by reviewers less familiar with the tool. This aspect needs to be further explored to determine the level of orientation required to use the tool appropriately. After finalizing the appraisal process, all three reviewers provided multiple suggestions for refining the draft dictionary to increase its clarity.

Table 5: Summary Kappa – three reviewers 13 studies, 18 criteria

Raters	K Value [SE], interpretation
R1 vs. R2	0.619 [0.056] Substantial agreement*
R1 vs. R3	0.592 [0.058] Moderate agreement*
R2 vs. R3	0.531 [0.061] Moderate agreement*

* Interpretation of Kappa values: no agreement: K < 0; slight agreement: 0.0 to 0.20; fair agreement: 0.21 to 0.40; moderate agreement: 0.41 to 0.60; substantial agreement: 0.61 to 0.80; perfect agreement: 0.81 to 1.00.
Abbreviations: K: interrater reliability (Kappa value); R: reviewer; SE: standard error.

Table 6: Summary Kappa – three reviewers versus their consensus scorings

Raters	K Value [SE], interpretation
R1 vs. consensus	0.806 [0.043] Substantial agreement*
R2 vs. consensus	0.806 [0.043] Substantial agreement*
R3 vs. consensus	0.552 [0.060] Moderate agreement*

* Interpretation of Kappa values: no agreement: K < 0; slight agreement: 0.0 to 0.20; fair agreement: 0.21 to 0.40; moderate agreement: 0.41 to 0.60; substantial agreement: 0.61 to 0.80; perfect agreement 0.81 to 1.00.
Abbreviations: K: interrater reliability (Kappa value); R: reviewer; SE: standard error.

Variation in rating the included studies

Variation across studies

Table 7 shows the rating of the 13 case series studies by the three reviewers. The mean values are calculated from the number of yes responses accumulated by each study when applying the checklist. The standard deviation values show which studies reached the lowest level of agreement among the reviewers. A low standard deviation value signals good agreement between reviewers whereas a high standard deviation signals poor agreement. Most of the studies with high levels of disagreement (e.g. studies 3, 6, 7, and 10) also had relatively low mean values, indicating that good, clearly reported studies contribute to the increase in agreement between reviewers. The span of variability across the 13 studies suggests that using the checklist enables researchers to differentiate between studies.

Table 7: Means and standard deviations of the rating for each study across the three reviewers, 18 criteria

Study	Mean*†	SD*†
1. Froud et al.[60]	0.722	0.192
2. Ryan et al.[61]	0.703	0.032
3. Lee et al.[59]	0.537	0.257
4. Hering et al.[62]	0.759	0.225
5. Hering et al.[63]	0.833	0.064
6. Hirshberg et al.[64]	0.648	0.225
7. Barshes et al.[65]	0.556	0.225
8. Shapiro et al.[66]	0.851	0.064
9. Maffi et al.[67]	0.796	0.096
10. Venturini et al.[68]	0.481	0.225
11. O'Connell et al.[69]	0.740	0.128
12. Poggioli et al.[71]	0.574	0.160
13. Keymeulen et al.[70]	0.870	0.096

*Considered all 18 criteria from the checklist; † calculated with Microsoft Office Excel 2003

Abbreviations: SD: standard deviation.

Variation across criteria

Table 8 highlights the extent of agreement among the three reviewers, who applied the 18-criteria checklist to the 13 reviewed studies. The mean values show that two criteria (criteria 1, aim/objective of the study reported, and 13, length of follow-up reported) were scored consistently across all the reviewers whereas criteria 3, 5, 6, and 18 were scored yes for only a few studies. Criterion 3 (characteristics of participants in the study described) and criterion 5 (participants recruited consecutively) were not met in 80% (mean value 0.103) and 90% (mean value 0.205) of the studies, respectively. Information about criteria 6 (participants entering the study at a similar point in the disease) and 18 (competing interests and sources of support reported) was available in only 38% (mean value 0.385) and 40% (mean value 0.410) of the studies examined.

The standard deviation values (Table 8) indicate the reviewers' level of agreement. A low standard deviation value for a criterion signals good agreement between reviewers for scoring the 13 studies whereas a high standard deviation signals poor agreement. For example, the reviewers showed perfect agreement in appraising the studies for criteria 1 and 13 whereas the highest variation among the raters was noted for the following criteria: 14 (loss to follow-up reported), 15 (provision of estimates of the random variability), 8 (additional interventions/co-interventions reported), 10 (relevant outcomes measured appropriately), and 12 (use of appropriate statistical tests to assess outcomes). The reasons for the poor agreement among reviewers for some of the criteria might include differences in interpretation of the relevant information in the study (e.g. criterion 8); variations in reviewer knowledge (e.g. criteria 12 and 15, which focus on statistical aspects) or understanding of the disease and the intervention of interest (e.g. criterion 8); reviewers overlooking relevant information reported in tables, figures, or boxes, or dispersed throughout the study (e.g. criteria 8, 14, and 15); or lack of clarity in the instructions provided in the draft dictionary (e.g. criteria 8, 10, and 15). This emphasizes the need for preliminary discussions and examination of the checklist criteria and instructions to increase consistency; to clarify some terms, such as "relevant" or "appropriate"; and to provide orientation to the clinical aspects of the disease and interventions.

Table 8: Means and standard deviations of ratings for each criterion across three reviewers, 13 studies

Criterion	Mean*†	SD*†	Criterion	Mean*†	SD*†
1	1.000	0.000	10	0.821	0.222
2	0.897	0.133	11	0.897	0.133
3	0.205	0.178	12	0.769	0.222
4	0.615	0.178	13	1.000	0.000
5	0.103	0.044	14	0.795	0.266
6	0.385	0.178	15	0.692	0.266
7	0.846	0.089	16	0.795	0.133
8	0.513	0.222	17	0.897	0.178
9	0.923	0.133	18	0.410	0.178

*Considered all 13 studies included in the pilot; † calculated with Microsoft Office Excel 2003.
Abbreviations: SD: standard deviation

Application of Quality Assessment Results

The Delphi panel did not develop a scoring system or guidance for the checklist when the pilot project was conducted. The number of yes responses based on reviewer consensus was counted for each of the 13 studies. The maximum number of yes responses for a study is 18 as each criterion was weighted equally. A study with 14 or more yes responses (≥ 70%) was considered to be of acceptable quality. Overall the quality rating of each of the studies was analyzed and the sentinel criteria that introduce risk of bias were identified.[58] These included consecutive cases, key outcomes measured before and after the intervention, and information provided on the loss to follow-up. The quality scores were not used as inclusion or exclusion criteria for the 13 case series studies.

The results from the pilot study indicate that it would be useful to provide a set of instructions and to guide reviewers on how to incorporate the quality appraisal results in the discussion and interpretation of research findings.

DISCUSSION

Comprehensiveness of the Checklist

The brief review of published checklists identified 36 checklists (Table 1) which aimed to appraise case series studies only (10 checklists) or studies of various designs, including case series (26 checklists) (Table C.1, Appendix C). These numbers indicate the paucity of checklists designed to appraise case series studies only; in the majority of studies the authors used substituted checklists developed for a different study design (i.e. randomized controlled trials or nonrandomized comparative studies) to solve this conundrum.

Only three publications (one including a checklist developed specifically for case series studies) provided more details on the processes undertaken to develop the checklist and the test results for reliability or validity. To capture a broader range of potential criteria, we included other publications that used or adapted an existing checklist in the review. Among them is the inventory composed of the 30 criteria[10] used by the Delphi panel as a basis for developing the 18-criteria checklist. The 30-criteria list was generated from five studies identified through a limited literature search. Although a comprehensive review of published checklists should have been conducted before initiating the Delphi process, a closer examination of the results from the brief review conducted after the Delphi process revealed that only a few criteria were missed in the initial 30-criteria list (reporting appropriate methods for recruitment of participants and defining objective methods to measure outcomes), indicating the comprehensiveness of the 30-criteria checklist. The panelists were content with the initial exhaustive list of 30-criteria and did not add new items in the first round of the Delphi process.

The criteria included in the 10 checklists developed for the appraisal of case series studies are summarized by frequency in Table 9. The far right column of Table 9 includes the 18-criteria voted for inclusion in the final checklist by the Delphi panel.

Table 9: Comparison of commonly used criteria found in the published checklists for case series studies only and the 18-criteria checklist

No.	Criterion	Number of studies that included the criterion (N = 10)	18-criteria checklist for CS (by Delphi panel)
1	Clear definition of the primary and secondary outcomes	8*	✓
2	Description of the eligibility criteria (inclusion and exclusion criteria)	8*	✓
3	Relevant outcomes reported	6*	✓ (1/2)†
4	Appropriate methods for recruitment of participants (e.g. adequate sample, relevant population)	5	–
5	Duration of follow-up reported and appropriate	5*	✓
6	Prospective study design	5*	–
7	Blind assessment of outcomes	5*	–
8	Participants recruited consecutively	4*	✓
9	Report of loss to follow-up	4*	✓
10	Description of the intervention	4*	✓
11	Description of the purpose, aim, or objectives of the study	4*	✓
12	Statistical tests appropriate/valid	3*	✓

13	Description of adverse events/side effects	3*	✓
14	Participants entering the study at a similar point in their disease progression	3*	✓
15	Description of co-intervention(s) received	3*	✓
16	Description of baseline characteristics of participants, such as age and gender	2*	✓
17	Objective methods to measure outcomes	2	✓ (1/2)†
18	Recruitment period clearly stated; same time	2*	–
19	Case series collected in more than one centre (multicentre study)	2*	✓
20	Outcomes measured before and after intervention	1*	✓
21	Study provides estimates of the random variability in the data analysis of relevant outcomes	1*	✓
22	Conclusions of the study supported by results	1*	✓
23	Both competing interests and sources of support for the study are reported	1*	✓

* Includes the broad checklist for quality appraisal of CS studies[10] used in the first round of the Delphi process; † part of one criterion in the 18-criteria checklist: "Were relevant outcomes appropriately measured with objective and/or subjective methods?"; ✓: yes, criterion is included in the 18-criteria checklist; –: no, criterion is not included in the 18-criteria checklist

Abbreviations: CS: case series; N: number of publications

As shown in Table 9, three criteria (i.e. description of appropriate methods for recruitment of participants, prospective design of the case series study, and blind assessment of outcomes) were included in 5 out of 10 checklists but not in the final checklist produced by the Delphi panel.

A prospective design allows baseline measurement and ensures that patient selection criteria, treatment protocol used, and outcome measures are predefined and standardized. Although this criterion was not included in the 18-criteria checklist, including other criteria about participant recruitment such as consecutive enrollment, defining objective methods of measuring outcomes, and providing measurements before and after the intervention might provide useful information about the methodological features and potential limitations/biases in selection, performance, and reporting the results of the study.

In one review published by NICE in 2005, the authors looked at the relationship between characteristics of case series studies and their outcomes.[2,3] They found little evidence to support the use of many of the criteria included in tools used for quality assessment of case series studies. The analyses were limited by poor reporting of methodological features. The authors did not find a relationship between study size and outcome, and a prospective approach was not associated with the outcome. Insufficient data were available to explore the aspects about consecutive recruitment or multi- versus single-centre studies. Mixed results were obtained for length of follow-up, independence of outcome measurement, and publication date. Findings from the NICE review should be interpreted with caution, however, because of several noted limitations, such as its narrow focus on surgical interventions, the inclusion of only a small number of cases in the studies examined, and the relatively limited number of studies in each group of case series examined for a specific intervention. The authors have called for further research into the determinants of quality for case series studies.

Table 9 shows four criteria (20 to 23) included in the original checklist of 30-criteria[10] but not included in any of the checklists identified in the brief review. One related to competing interests

and sources of financial support is identified as important in the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, which established a checklist of 22 criteria.[72,73] The STROBE recommendations are currently limited to three major study designs: cohort, case-control, and cross sectional. If the approach proves applicable and improves reporting, its application to other designs such as case series can be considered in future updates of the STROBE criteria.[72] In a recent review[74] the authors evaluated the quality of reporting of interventional case series studies and randomized controlled trials (RCTs) published between 2005 and 2007 in the field of ophthalmology. They used two different reporting standards, the STROBE checklist for reporting case series studies and the Consolidated Standards of Reporting Trials (CONSORT) checklist. The authors found that the small case series interventional studies had an average reporting score lower than the RCTs; however, in some instances the case series studies received higher scores than RCTs. The authors also stated that although good reporting is not a direct measure of the quality of a study, it allows a reader to assess the validity and applicability of the study's findings.[74]

The review of published checklists for the appraisal of case series studies was found helpful for informing about the possibility of adding or excluding some of the criteria from the checklist developed by the Delphi panel and contributed to the verification of its completeness in comparison to other published checklists.

Feasibility and Usefulness of the Checklist

Some of the published quality appraisal checklists found in the literature search include scales in which each criterion has a numeric score attached to it with an overall summary score (Table 1). Although some authors recommend the use of scales, others consider them potentially misleading.[5] There are still uncertainties about the relationship between methodological features and validity and how criteria scores should be summed into a single measure of study quality.[3] The Delphi panel did not attempt to develop a scale or any guidance for the final 18-criteria checklist. Checklist users may decide to define a cut-off point to separate the "high-quality" studies (studies that meet a certain number or percentage of criteria) from the "low-quality" ones. They might also identify some criteria from the checklist which are relevant to a specific condition or technology, and they can focus more on discussing the outcomes from the studies that meet those selected criteria.

The 18-criteria checklist was piloted by three reviewers on 13 case series studies selected for inclusion in an HTA review of islet transplantation for type 1diabetes. Two reviewers customized the draft dictionary and discussed the checklist before commencing the evaluation. This process increased their understanding of the requirements of the checklist and probability of achieving consensus. The third reviewer was invited at a later stage with the intent to gain an understanding of how the checklist works by an assessor at arm's length from the process.

In general, the draft customized dictionary was found useful and easy to use; however, the dictionary for some criteria was still vague enough to necessitate iterative discussions and several modifications to improve the clarity. Some criteria were difficult to score and encountered more disagreements among reviewers. An example is the criterion: entering the study at a similar point in the disease. The complexity of the specific clinical condition investigated might have been the reason for difficulties in scoring this criterion, and this limitation may be overcome by including a clinical expert from the investigated area in the appraisal process. Another example is the criterion: were the statistical tests used to assess the relevant outcomes appropriate? In this case the suggestion was to consider the involvement of a statistician in the appraisal process.

Suggestions for improvements to the dictionary were collected from the reviewers involved in the pilot and two other reviewers involved in the Delphi panel (Appendix K). These included:

- adding a third choice option to the dictionary (i.e. partially responded or unclear) to increase their clarity,
- refinement of the wording of some of the criteria,
- inclusion of supplementary notes for reviewers, and
- inclusion of two new criteria commonly used in other case series studies quality appraisal checklists (i.e. "Was the study conducted prospectively?" and "Were the main outcomes assessed blind to/independent of intervention status?").

The dictionary may need to be customized for each review, and reviewers involved in the appraisal process should determine which criteria are crucial for a specific situation (condition and technology).

The exact time required for applying the checklist to each of the 13 studies was not recorded. The estimated time varied across the studies, ranging from 30 minutes to 2 hours. Furthermore, because of the variation in the lengths of the publications and clarity of presentation and reporting, the time spent on each study was not necessary related to increased reviewer familiarity with the tool.

The participants in the pilot study generally felt that the checklist needs further improvements before its widespread use. It appears to be too lengthy for the methodological appraisal of a case series study, which, by definition, represents the weakest study design to determine treatment benefits. Through the analyses of convergence of individual judgments and agreements during the three-round process, it was possible to decrease the number of criteria from 30 to 18. The expertise and experience of panelists (i.e. specialized in HTA) might have influenced the selection process of criteria. An attempt to shorten the checklist was made in round 3; however, the panelists excluded only one criterion.

The majority of the criteria in the checklist focus on the reporting aspects of the study whereas a few criteria examine how the study was conducted or executed (e.g. design of the study as a multicentre trial, inclusion of consecutive cases, clear measurements of baseline characteristics and outcomes, or reporting of loss to follow-up). The rating results of these key criteria should be highlighted and incorporated in the interpretation of the research evidence.

Interrater Reliability

The interrater reliability measured by Kappa coefficients demonstrated moderate to substantial agreement among the three reviewers involved in the pilot study. Some of the differences in the appraisal may have been related to the reviewer or the difficulty of the subject.

Attempts were made to calculate Kappa coefficient values for individual criteria. Unfortunately, the pilot included an insufficient number of case series studies, and variation in the ratings among the studies was not substantial enough to allow precise calculations of the Kappa measurements. More reviewers should be involved in a future test of the tool; in particular, a comparison should be made of the results from trained and untrained reviewers to identify if other factors, such as levels of personal skills or knowledge, contribute to the discrepancies in the rating results.

Study Limitations

Several methodological limitations are noteworthy.

The initial 30-item list was generated from five studies identified through a limited literature search. A comprehensive search was subsequently conducted, locating 36 studies that developed or used a quality assessment checklist for case series or studies of various designs, including case series studies. Although the comprehensive search should have been conducted before the Delphi process, a close review of the additional studies revealed that no important items were missed in the 30-criteria list.

The judgments involved in creating the checklist are those of a self-selected group of seven HTA professionals who rated each criterion based on their personal perception about its importance. Hence the checklist may not reflect the criteria seen to be crucial for assessing methodological quality by reviewers outside of the HTA field. Another limitation in selecting the panelists is the absence of panel member with expertise in conducting primary case series studies.

The small number of raters (i.e. three) and included case series studies (i.e. 13) made it difficult to calculate Kappa values precisely. A significantly larger number of case series studies should be assessed by more raters in order to conduct a proper evaluation of the reliability and validity of the checklist.

FUTURE IMPROVEMENTS OF THE CHECKLIST

The pilot study provided some insights into how criteria included in the checklist work, and, further, the information may be used in judging the inclusion or exclusion of some of the criteria in the checklist. For example, if we expect that one criterion will always be present in all the studies examined, we may decide to drop that criterion to reduce the number of criteria from the checklist. Nevertheless, more analyses, including a factor analysis, would be needed before a decision can be made to eliminate criteria.

A preliminary factor analysis was conducted in an attempt to identify redundant items and to shorten the checklist (Appendix L). However, due to the small number of case series studies included in the pilot, the findings from the preliminary factor analysis failed to identify any items that should be removed from the checklist.

As an extension of this pilot, quality rating of a larger number of studies using this tool will be performed in the future, and a more comprehensive factor analysis will be conducted in the hope of further shortening the checklist. Furthermore, a new study phase was planned that will investigate the dimensionalities and construct validity of the checklist to evaluate the concurrent validity and to assess the reliability properties of the instrument. These investigations will help to clarify the role of the checklist as a useful tool for the assessment of case series studies.

CONCLUSION

An 18-criteria checklist was developed by a panel of seven HTA researchers using a modified Delphi method. Although a search of the published case series quality assessment checklists prior to the Delphi process would have been useful to highlight prevalence and importance of some criteria against others, the brief review conducted after finalization of the checklist indicated the comprehensiveness of the inventory checklist used by the Delphi panel.

Three HTA reviewers piloted the checklist to assess the methodological quality of 13 case series studies on islet transplantation for the treatment of type 1 diabetes. This exercise provided first-hand experience with the checklist and identified several places for further improvement. The researchers found the checklist to be a useful tool, and, with some further adaptations and modifications, it can be used in future HTA work.

Modifications are required for both the checklist and the dictionary. The criteria within the checklist need to be reduced in number and their clarity improved in order to make the tool more user friendly. Of the many suggestions for modifications, two additional criteria on prospective study design and blind assessment of outcomes might need to be reconsidered. As part of the evaluation of the tool, a factor analysis is planned in the near future, followed by additional testing of its validity and reliability by a larger group of reviewers.

Given the paucity of quality assessment tools specifically developed for case series studies, the 18-criteria checklist serves as a good starting point to examine the quality of case series studies. However, HTA researchers might find that what constitute acceptable assessment criteria vary under different situations and that additional criteria specific to the subject area are often required. For this reason, although a useful tool, the resultant checklist should not be viewed as a definitive solution; instead, it should be viewed and used as a guide only.

APPENDIX A: LITERATURE SEARCH FOR EXISTING QUALITY ASSESSMENT TOOLS FOR CASE SERIES STUDIES

The literature search was conducted by the IHE Research Librarians (Patricia Chatterley and Liz Dennett) for articles published between 1998 and June 2011. The search was developed and carried out prior to the study selection process. In addition to the strategy outlined below, reference lists of retrieved articles were reviewed for potential studies.

Table A.1: Search strategy

Database	Edition or date searched	Search Terms[††]
Databases		
The Cochrane Library www.thecochrane library.com	1998 to June 28, 2011	checklist* or check-list* or scale or scales or tool or tools or instrument* or keys ***in*** *Title, Abstract or Keywords* AND quality or validity or reliability *in Title, Abstract or Keywords* AND "case series" or noncomparative or "single arm" or "single group" or "observational study" or "observational studies" *in Title, Abstract or Keywords*, from 1998 to 2011 "critical* treatmen* *in Title, Abstract or Keywords* and "case series" or noncomparative or "single arm" or "single group" or "observational study" or "observational studies" *in Title, Abstract or Keywords*, from 1998 to 2011
MEDLINE	1998 to June 29, 2011	1. (((check adj list$) or check-list$ or checklist$ or scale or scales or tool or tools or key or keys or instrument or instruments or form or forms) adj3 (quality or validity or validat$ or assess$ or evaluat$ or treatmen$ or measure$ or review$ or analy$ or judg$) adj5 (study or studies or paper$ or article$ or literature or report$ or research)).mp. 2. ((assess$ or treatmen$ or judg$ or measure$ or analy$ or evaluat$) adj3 quality adj3 (research or literature or report$ or paper$ or study or studies or article$)).ti. 3. (critical$ adj treatmen$ adj3 (research or literature or report$ or paper$ or study or studies or article$)).ti. 4. or/1-3 5. (observational or case or case series or time series or noncomparative or single arm or single group).mp. 6. (nonrandom$ or non-random$).mp. 7. 5 or 6 8. 4 and 7 9. exp Clinical Trials as Topic/ 10. 8 and 9 11. ((observational or case or case series or time series or noncomparative or single arm or single group) adj (study or studies or data)).mp. 12. 4 and (6 or 11) 13. (quality or reliability or validity).mp. 14. 12 and 13 15. 10 or 14 16. limit 15 to yr="1998 – 2011"

CRD Databases (DARE, HTA & NHS EED)	1998 to June 29, 2011	# 1 checklist* NEAR quality # 2 check-list* NEAR quality # 3 scale* NEAR quality # 4 tool* NEAR quality # 5 instrument* NEAR quality # 6 keys NEAR quality # 7 checklist* NEAR validity # 8 check-list* NEAR validity # 9 scale* NEAR validity # 10 tool* NEAR validity # 11 instrument* NEAR validity # 12 instrument* NEAR reliability # 13 checklist* NEAR reliability # 14 check-list* NEAR reliability # 15 scale* NEAR reliability # 16 tool* NEAR reliability # 17 #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16 # 18 “case series” # 19 #17 AND #18 # 20 #17 AND #18 RESTRICT YR 1998 2008 # 21 critical* AND appraisal* # 22 #18 AND #21 # 23 #19 OR #22 # 24 #19 OR #22 RESTRICT YR 1998 2008
EMBASE –Ovid platform (Licensed resource)	1998 to 2011 Week 25	1. (((check adj list$) or check-list$ or checklist$ or scale or scales or tool or tools or key or keys or instrument or instruments or form or forms) adj3 (quality or validity or validat$ or assess$ or evaluat$ or treatmen$ or measure$ or review$ or analy$ or judg$) adj5 (study or studies or paper$ or article$ or literature or report$ or research)).mp. 2. ((assess$ or treatmen$ or judg$ or measure$ or analy$ or evaluat$) adj3 quality adj3 (research or literature or report$ or paper$ or study or studies or article$)).ti. 3. (critical$ adj treatmen$ adj3 (research or literature or report$ or paper$ or study or studies or article$)).ti. 4. or/1-3 5. case study/ 6. (((observational or case or case series or time series or noncomparative or single arm or single group) adj2 (study or studies or data)) or case series).mp. 7. (nonrandom$ or non-random$).mp. 8. or/5-7 9. 4 and 8 10. (quality or reliability or validity).mp.

		11. 9 and 10 12. limit 11 to yr="1998 – 2008"

Note:

The characters "††", "*", "# ", and "?" are truncation characters that retrieve all possible suffix variations of the root word; e.g. surg* retrieves surgery, surgical, surgeon, etc.

Semicolons (;) are used to separate search terms that were searched separately.

APPENDIX B: EXCLUDED PUBLICATIONS

Table B.1: Excluded publications

Main reason for exclusion: **Not clear if the checklist was designed for case series studies or studies of various designs**
Met R, Van Lienden KP, Koelemay MJW, Bipat S, Legemate DA, Reekers JA. Subintimal angioplasty for peripheral arterial occlusive disease: a systematic review. *Cardiovascular and Interventional Radiology2008;*31(4):687-97.
Main reason for exclusion: **Used same checklists as studies included in Table C.1, Appendix C, for quality appraisal of case series studies only**
Bala MM, Riemsma RP, Nixon J, Kleijnen J. Systematic review of the (cost-)effectiveness of spinal cord stimulation for people with failed back surgery syndrome. *Clinical Journal of Pain* 2008;24(9):741-56 (same as CRD's Guidance, 2001).
Bilney B, Morris ME, Perry A. Effectiveness of physiotherapy, occupational therapy, and speech pathology for people with Huntington's disease: a systematic review. *Neurorehabilitation & Neural Repair* 2003;17(1):12-24 (same as CRD's Guidance, 2001).
Christie A, Dagfinrud H, Engen MK, Flaatten HI, Ringen OH, Hagen KB. Surgical interventions for the rheumatoid shoulder. *Cochrane Database of Systematic Reviews* 2010, (1):CD006188 (same as CRD's Guidance, 2001).
Lewis R, Bagnall AM, Forbes C, Shirran E, Duffy S, Kleijnen J, et al. The clinical effectiveness of trastuzumab for breast cancer: a systematic review. *Health Technology Assessment* 2002;6(13):1-71 (same as CRD's Guidance, 2001).
Nicholson T, Milne R. *Pallidotomy, thalamotomy and deep brain stimulation for severe Parkinson's disease.* Report: 63, 1999 (used old version CRD guidance 1996).
Medical Services Advisory Committee. *Placement of artificial bowel sphincters in the management of faecal incontinence: assessment report.* Canberra, Australia: MSAC; 2003 May (same as Young et al. 1999).
Wake B, Hyde C, Bryan S, Barton P, Song F, Fry-Smith A, Davenport C. Rituximab as third-line treatment for refractory or recurrent Stage III or IV follicular non-Hodgkin's lymphoma: a systematic review and economic evaluation. *Health Technology Assessment* 2002;6(3):1-85 (same as Young et al. 1999).
Main reason for exclusion: **Used same checklists as studies included in Table C.1, Appendix C, for quality appraisal of case series studies and studies of other design**
Berney S. B. The acute respiratory management of cervical spinal cord injury in the first 6 weeks after injury: a systematic review. *Spinal Cord* 2011;49(1):17-29 (same as Wells et al – Newcastle-Ottawa Quality Assessment Scale).
Cody J, Wyness L, Wallace S, Glazener C, Kilonzo M, Stearns S, et al. Systematic review of the clinical effectiveness of tension-free vaginal tape for treatment of urinary stress incontinence. *Health Technology Assessment* 2003;7(21):1-202 (same as Downs & Black, 1998).
Darrah J, Watkins B, Chen L, Bonin C. *Effects of conductive education intervention for children with a diagnosis of cerebral palsy: an AACPDM evidence report.* Report: 34, 2003 (same as AACPDM methodology, 2008).
Dodd KJ, Taylor NF, Damiano DL. A systematic review of the effectiveness of strength-training programs for people with cerebral palsy. *Archives of Physical Medicine and Rehabilitation 2002;*83(8):1157-64 (same as PEDro Scale, 1998).
Franzetti F, Antonelli M, Bassetti M, Blasi F, Langer M, Scaglione F, et al. Consensus document on controversial issues for the treatment of hospital-associated pneumonia. *International Journal of Infectious Diseases* 2010;14 Suppl 4:S55-S65 (same as Wells et al – Newcastle-Ottawa Quality Assessment Scale).
Gibson K, Growse A, Korda L, Wray E, MacDermid JC. The effectiveness of rehabilitation for nonoperative management of shoulder instability: a systematic review. *Journal of Hand Therapy* 2004;17(2):229-42 (same as MacDermid, 2003).
Greenburg DL, Lettieri CJ, Eliasson AH. Effects of surgical weight loss on measures of obstructive sleep apnea: a meta-analysis. *American Journal of Medicine* 2009;122(6):535-42 (modified checklist of Downs & Black, 1998).
Harris SR, Roxborough L. Efficacy and effectiveness of physical therapy in enhancing postural control in children with cerebral palsy. *Neural Plasticity* 2005;12(2-3):229-43 (same as AACPDM methodology, 2008).
Hiremath S, Holden RM, Fergusson D, Zimmerman DL. Antiplatelet medications in hemodialysis patients: a systematic review of bleeding rates. *Clinical Journal of The American Society of Nephrology* 2009;4(8):1347-55 (same as Wells et al – Newcastle-Ottawa Quality Assessment Scale).

Jamula E, Anderson J, Douketis JD. Safety of continuing warfarin therapy during cataract surgery: a systematic review and meta-analysis. *Thrombosis Research* 2009;124(3):292-9 (same as Wells et al – Newcastle-Ottawa Quality Assessment Scale).
Krassioukov A, Eng JJ, Warburton DE, Teasell R. A systematic review of the management of orthostatic hypotension after spinal cord injury. *Archives of Physical Medicine and Rehabilitation* 2009;90(5):876-85 (same as Downs & Black, 1998).
Leone S, Borre S, Monforte A, Mordente G, Petrosillo N, Signore A, et al. Consensus document on controversial issues in the diagnosis and treatment of prosthetic joint infections. *International Journal of Infectious Diseases* 2010;14 Suppl 4:S67-S77 (same as Wells et al – Newcastle-Ottawa Quality Assessment Scale).
Malcomson KS, Dunwoody L, Lowe-Strong AS. Psychosocial interventions in people with multiple sclerosis: a review. *Journal of Neurology* 2007;254(1):1-13 (same as Downs & Black, 1998).
Muller M, Tsui D, Schnurr R, Biddulph-Deisroth L, Hard J, MacDermid JC. Effectiveness of hand therapy interventions in primary management of carpal tunnel syndrome: a systematic review. *Journal of Hand Therapy* 2004;17(2):210-228 (same as MacDermid, 2003).
O'Brien MA, Villias-Keever M, Robinson P, Skye A, Gafni A, Brouwers M, et al. *Impact of cancer-related decision aids.* Report: 386, 2002 (same as Downs & Black, 1998).
Scheer MG, Sloots CE, van der Wilt GJ, Ruers TJ. Management of patients with asymptomatic colorectal cancer and synchronous irresectable metastases. *Annals of Oncology* 2008;19(11):1829-35 (same as Downs & Black, 1998).
Sheel AW, Reid WD, Townson AF, Ayas NT, Konnyu KJ, Spinal Cord Rehabilitation Evidence Research Team. Effects of exercise training and inspiratory muscle training in spinal cord injury: a systematic review. *Journal of Spinal Cord Medicine* 2008;31(5):500-8 (same as Downs & Black, 1998).
Smith TO, Leigh D. Outcomes following trochleoplasty for patellar instability with trochlear dysplasia: a systematic review. *European Journal of Orthopaedic Surgery and Traumatology* 2008;18(6):425-33 (same as Smith et al., 2009).
Stasi R, Sarpatwari A, Segal JB, Osborn J, Evangelista ML, Cooper N, et al. Effects of eradication of *Helicobacter pylori* infection in patients with immune thrombocytopenic purpura: a systematic review. *Blood* 2009;113(6):1231-40 (same as Wells et al. – Newcastle-Ottawa Quality Assessment Scale).
Stuber KJ, Smith DL. Chiropractic treatment of pregnancy-related low back pain: a systematic review of the evidence. *Journal of Manipulative & Physiological Therapeutics* 2008;31(6):447-54 (same as Downs & Black, 1998).
Teasell RW. Venous thromboembolism after spinal cord injury. *Archives of Physical Medicine and Rehabilitation* 2009;90(2):232-45 (same as Downs & Black, 1998).
Tilney HS, Tekkis PP. Extending the horizons of restorative rectal surgery: intersphincteric resection for low rectal cancer. *Colorectal Disease* 2008;10(1):3-15 (same as Slim et al. 2003).
Verschuren O, Ketelaar M, Takken T, Helders PJ, Gorter JW. Exercise programs for children with cerebral palsy: a systematic review of the literature. *American Journal of Physical Medicine and Rehabilitation* 2008;87(5):404-17 (same as PEDro Scale, 1998).
Wessel J. The effectiveness of hand exercises for persons with rheumatoid arthritis: a systematic review. *Journal of Hand Therapy* 2004;17(2):174-80 (same as MacDermid, 2003).
Main reason for exclusion: **Did not include a quality appraisal checklist; graded studies based on study design**
De Rango P, Cao P, Parlani G, Verzini F, Brambilla D. Outcome after endografting in small and large abdominal aortic aneurysms: a metanalysis. *European Journal of Vascular and Endovascular Surgery* 2008;35(2):162-72.
Liu SL, Lebrun CM. Effect of oral contraceptives and hormone replacement therapy on bone mineral density in premenopausal and perimenopausal women: a systematic review. *British Journal of Sports Medicine* 2006;40(1):11-24.
Manterola C, Pineda V, Vial M, Losada H, Munoz S. Surgery for morbid obesity: selection of operation based on evidence from literature review. *Obesity Surgery* 2005;15(1):106-13.
Ostaszkiewicz J, Ski C, Hornby L. Does successful treatment of constipation or faecal impaction resolve lower urinary tract symptoms: a structured review of the literature − systematic review. *Australian and New Zealand Continence Journal* 2005;11(3):70, 72, 74-75, 77-80.
Soldani F, Ghaemi SN, Baldessarini RJ. Research reports on treatments for bipolar disorder: preliminary assessment of methodological quality. *Acta Psychiatrica Scandinavica* 2005;112(1):72-4.

Main reason for exclusion: **Did not appraise the quality of case series studies**
Bagnall AM, Jones L, Richardson G, Duffy S, Riemsma R. Effectiveness and cost-effectiveness of acute hospital-based spinal cord injuries services: systematic review. *Health Technology Assessment (Winchester, England)* 2003;7(19):iii-92.
Choi PT, Galinski SE, Takeuchi L, Lucas S, Tamayo C, Jadad AR. PDPH is a common complication of neuraxial blockade in parturients: a meta-analysis of obstetrical studies. *Canadian Journal of Anaesthesia* 2003;50(5):460-69.
Espallargues M, Pons JM. Efficacy and safety of viscosupplementation with Hylan G-F 20 for the treatment of knee osteoarthritis: a systematic review. *International Journal of Technology Assessment in Health Care* 2003;19(1):41-56.
Jampel HD, Friedman DS, Lubomski LH, Kempen JH, Quigley H, Congdon N, et al. Methodologic rigor of clinical trials on surgical management of eyes with coexisting cataract and glaucoma. *Ophthalmology* 2002;109(10):1892-1901.
Ling E, Arellano R. Systematic overview of the evidence supporting the use of cerebrospinal fluid drainage in thoracoabdominal aneurysm surgery for prevention of paraplegia. *Anesthesiology* 2000;93(4):1115-22.
Ross SD, DiGeorge A, Connelly JE, Whiting GW, McDonnell N. Safety of GM-CSF in patients with AIDS: a review of the literature. *Pharmacotherapy* 1998;18(6):1290-97.
Tilney HS, Constantinides V, Ioannides AS, Tekkis PP, A. W. Darzi, Haddad MJ. Pouch-anal anastomosis vs straight ileoanal anastomosis in pediatric patients: a meta-analysis. *Journal of Pediatric Surgery* 2006;41(11):1799-1808.
Verstraaten J, Kuijpers-Jagtman AM, Mommaerts MY, Berge SJ, Nada RM, Schols JG, et al. A systematic review of the effects of bone-borne surgical assisted rapid maxillary expansion. *Journal of Cranio-Maxillo-Facial Surgery* 2010;38(3):166-74.
Wechter ME, Wu JM, Marzano D, Haefner H. Management of Bartholin duct cysts and abscesses: a systematic review. *Obstetrical and Gynecological Survey* 2009;64(6):395-404.
Main reason for exclusion: **Did not include case series studies in the appraisal**
Armenteros JL, Davies M. Antipsychotics in early onset schizophrenia: systematic review and meta-analysis. *European Child and Adolescent Psychiatry* 2006;15(3):141-48.
Dierick-van Daele ATM, Spreeuwenberg C, Derckx EWCC, Metsemakers JFM, Vrijhoef BJM. Critical appraisal of the literature on economic evaluations of substitution of skills between professionals: a systematic literature review. *Journal of Evaluation in Clinical Practice* 2008;1(4):481-92.
Gwadry-Sridhar FH. A framework for planning and critiquing medication compliance and persistence research using prospective study designs. *Clinical Therapeutics* 2009;3(2):421-35.
Hillingso JG, Wille-Jorgensen P. Staged or simultaneous resection of synchronous liver metastases from colorectal cancer–a systematic review. *Colorectal Disease* 2009;11(1):3-10.
Hiremath S, Holden RM, Fergusson D, Zimmerman DL. Antiplatelet medications in hemodialysis patients: a systematic review of bleeding rates. *Clinical Journal of The American Society of Nephrology* 2009;4(8):1347-55.
Huang J, Barbera L, Brouwers M, Browman G, Mackillop WJ. Does delay in starting treatment affect the outcomes of radiotherapy: a systematic review. *Journal of Clinical Oncology* 2003;21(3):555-63.
Kardamanidis K, Martiniuk A, Ivers RQ, Stevenson MR, Thistlethwaite K. Motorcycle rider training for the prevention of road traffic crashes. *Cochrane Database of Systematic Reviews* 2010;(10):CD005240.
Langham S, Langham J, Goertz HP, Ratcliffe M. Large-scale, prospective, observational studies in patients with psoriasis and psoriatic arthritis: a systematic and critical review. *BMC Medical Research Methodology* 2011;11:32.
Lanting B, MacDermid J, Drosdowech D, Faber KJ. Proximal humeral fractures: a systematic review of treatment modalities. *Journal of Shoulder and Elbow Surgery* 2008;17(1):42-54.
Liu T, Li L, Korantzopoulos P, Liu E, Li G. Statin use and development of atrial fibrillation: a systematic review and meta-analysis of randomized clinical trials and observational studies. *International Journal of Cardiology* 2008;126(2):160-70.
Meade MO, Herridge MS. An evidence-based approach to acute respiratory distress syndrome. *Respiratory Care* 2001;46(12):1368-79.
Pan A, Cauda R, Concia E, Esposito S, Sganga G, Stefani S, et al. Consensus document on controversial issues in the treatment of complicated skin and skin-structure infections. *International Journal of Infectious Diseases* 2010;14 Suppl 4:S39-S53.

Ross SD, Allen IE, Harrison KJ, Kvasz M, Connelly J, Sheinhait IA. *Systematic review of the literature regarding the diagnosis of sleep apnea.* Report: 154, 1999.
Vancampfort D, Knapen J, De HM, van WR, Deckx S, Maurissen K, et al. Cardiometabolic effects of physical activity interventions for people with schizophrenia. *Physical Therapy Reviews* 2009;14(6):388-98.
Wilson s, Maddison t, Roberts L, Greenfield S, Singh S. Systematic review: the effectiveness of hypnotherapy in the management of irritable bowel syndrome. *Alimentary Pharmacology and Therapeutics* 2006;24(5):769-80.
Walburn J, Gray R, Gournay K, Quraishi S, David AS. Systematic review of patient and nurse attitudes to depot antipsychotic medication. *British Journal of Psychiatry* 2001;179(4):300-7.
Main reason for exclusion: **Did not include a quality appraisal checklist; not clear if included case series studies**
Leichsenring F, Rabung S. Effectiveness of long-term psychodynamic psychotherapy: a meta-analysis. *JAMA* 2008;300(13):1551-65.
Price D, Jefferson T, Demicheli V. Methodological issues arising from systematic reviews of the evidence of safety of vaccines. *Vaccine* 2004;22(15-16):2080-84.
Main reason for exclusion: **Not clear which modifications of the checklist were made to appraise case series studies**
Schabrun SM, Hillier S. Evidence for the retraining of sensation after stroke: a systematic review. *Clinical Rehabilitation* 2009;23(1):27-39. (Modified checklist: Law M et al. Critical review form – quantitative studies. McMasters University Occupational Therapy Evidence-Based Practice Research Group, 1998.)
Main reason for exclusion: **Background documents, review of methodological aspects**
Bornhoft G, Maxion-Bergemann S, Wolf U, Kienle GS, Michalsen A, Vollmar HC, et al. Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity. *BMC Medical Research Methodology* 2006;6:56.
Margetts BM. Evidence-based nutrition: review of nutritional epidemiological studies. *South African Journal of Clinical Nutrition* 2002;15(3):68-73.
Ravaud P, Boutron I. Primer: assessing the efficacy and safety of nonpharmacologic treatments for chronic rheumatic diseases. *Nature Clinical Practice Rheumatology* 2006;2(6):313-19.
Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. *American Journal of Public Health* 2004;94(3):400-5.
Main reason for exclusion: **Checklist developed for specific condition or a specific intervention (e.g. diagnostic tests)**
Althaus F. Effectiveness of interventions targeting frequent users of emergency departments: a systematic review. *Journal of General Internal Medicine* 2010; Conference(var.pagings):S266-S267.
Everett CR, Shah R, Sehgal VN, Kenzie-Brown AM. A systematic review of diagnostic utility of selective nerve root blocks. *Pain Physician* 2005;8(2):225-33.
Fernandez R, Griffiths R, Halcomb E, Chow J. *The infection control management of MRSA within the acute care hospital.* Report: 1-81, 2002.
Gray M, Gold L, Burls A, Elley K. *The effectiveness of toluidine blue dye as an adjunct to oral cancer screening in general dental practice.* Report: 40, 2000.
Kon EV. Matrix-assisted autologous chondrocyte transplantation for the repair of cartilage defects of the knee: systematic clinical data review and study quality analysis. *The American Journal of Sports Medicine* 2009;37 Suppl 1:56S-66S.
Mellegers MA, Furlan AD, Mailis A. Gabapentin for neuropathic pain: systematic review of controlled and uncontrolled literature. *Clinical Journal of Pain* 2001;17(4):284-95.
McMillan D, Lee R. A systematic review of behavioral experiments vs exposure alone in the treatment of anxiety disorders: a case of exposure while wearing the emperor's new clothes? *Clinical Psychology Review* 2010;30(5):467-78.
Reuchlin-Vroklage LM, Bierma-Zeinstra S, Benninga MA, Berger MY. Diagnostic value of abdominal radiography in constipated children: a systematic review. *Archives of Pediatrics and Adolescent Medicine* 2005;159(7):671-78.

Abbreviations: AACPDM: American Academy for Cerebral Palsy and Development Medicine; CRD: Centre for Reviews and Dissemination; MSAC: Medical Services Advisory Committee

APPENDIX C: RESULTS – CRITICAL APPRAISAL TOOLS FOR CASE SERIES STUDIES

Table C.1: Methods used to develop the checklists and internal validation of the published tools

Study[a]	Source of the tool	Detailed method used to develop the checklist (criteria selection, process)	Interrater reliability reported	Instructions provided for scoring	Time for completion
Cauchi et al.[6] 2008	Adapted from other sources[7,21]	Independent review body[†]	No	No	Not stated
Chipchase et al.[15] 2009	Based on[75,76]	Expert researcher conducted literature search and suggested the tool	No	No	Not stated
CRD's Guidance[7] 2001*	Adapted from other checklists[77-79]	Proposed quality criteria	No	No	Not stated
Huisstede et al.[8] 2008	Adapted/modified from[80-82]	Not stated	No	Yes	Not stated
McCrory et al.[9] 2001*	Not stated	Not stated	No	No	Not stated
Moga & Harstall[10] 2006*	Adapted from[2,11,21,55,56]	Not stated	Yes, K = 0.58 (good)	Yes	Not stated
National Collaborating Centre for Acute Care[11] 2003*	Not stated	Not stated	No	No	Not stated
Taylor et al.[12] 2005	Adapted from[83,84]	Not stated	No	No	Not stated
Yang et al.[13] 2009	Experts	Panel of experts generated initial criteria; modified Delphi technique – rated importance of criteria; pretested; refinement of the instrument	Cronbach's α between 0.80 and 0.85 Intra-class correlations 0.904 (high)	Yes	≤15 minutes per paper
Young et al.[14] 1999*	Not stated	Not stated	No	Yes	Not stated
AACPDM methodology [16] 2008* Not designed for case series	Adapted, source(s) not stated	Subgroup committee (five experts). Review literature on conducting/quality scoring systems	No	Yes	Not stated
Bryant et al.[17] 2002*	Modified from[44]	Not stated	No	Yes	Not stated
Deenadayalan et al.[40] 2010	Modified from [85]	Not stated	No	No	Not stated
de Kleuver et al.[19] 2003	Adapted from[86]	Two reviewers regrouped criteria and added five more questions	No	No	Not stated
Des Jarlais et al.[20] 2004	Not stated	Not stated	No	No	Not stated

Downs & Black[21] 1998	Developed from other checklists used for the assessment of RCTs[45-51]	Pilot test followed by revisions; measurement of internal consistency, test-retest reliability, interrater reliability, face and criterion validity	r = 0.75 (good)	Yes	Average: 20 to 25 minutes per paper; Range: 10 to 45 minutes per paper
Green et al.[22] 1999*	Not stated	Not stated	No	Yes[87]	Not stated
Haines et al.[23] 1999	Adapted from[88]	Not stated	No	No	Not stated
Hardy et al.[24] 2001*	Not stated	Not stated	No	No	Not stated
Hayashi et al.[25] 2003	Adapted[89-93]	Calibration of quality assessment using a sample paper prior to assessment; intra-rater reliability	Good agreement (in almost 80% of the criteria assessed)	No	Not stated
Helm et al.[37] 2009	Adapted and modified from[55]	Not stated	No	No	Not stated
Jongerius et al.[26] 2003	Not stated	Not stated	No	No	Not stated
MacDermid[27] 2003	Not stated	Not stated	No	Yes	Not stated
Merlin et al.[28] 2001*	Not stated	Not stated	No	No	Not stated
Mortenson & Eng[29] 2003	Modified version[94]	Not stated	No	No	Not stated
Nichol et al.[18] 1999‡	Modified from Spitzer et al.[44]§	Eliminated criteria which did not address systematic bias and consistency; pretest instrument§	W = 0.64 r = 0.89 (95% CI 0.73 to 0.93) §	Yes§	Approx. 30 minutes per article/ reviewer§
Oremus et al.[30] 2008*	Adapted from[95,96]	Not stated	No	No	Not stated
Overend et al.[31] 2001	Developed by the authors	Critical appraisal form piloted to refine the criteria and information required	No	No	Not stated
PEDro Scale[32] 1998	Not stated	Not stated	No	Yes	Not stated
Reiman et al.[38] 2009	Used the Maastricht-Amsterdam criteria list[97]	Not stated	No	No	Not stated
Slim et al.[33] 2003	Not stated	Review literature; selected criteria; score the ability of each criterion to assess quality; pretest (measure inter-reviewer reliability, test-retest reliability, internal consistency); refine one criterion	Yes, K measured for each criterion	Yes	Not stated
Smith et al.[39] 2009	CASP[98]	Not stated	No	No	Not stated
Stead et al.[34] 2008	Wells et al.[41]	Not stated	Yes[41] (not stated)	Yes	Not stated
Steward et al.[35] 1999*	Not stated	Not stated	No	No	Not stated

Thomas et al.[36] 2006	Based on guidelines suggested by[22,86]	Not stated	No	Yes	Not stated
Wells et al.[41] 2000	Not stated	Not stated	Not reported	Yes	Not stated

[a] Studies are presented in alphabetical order by lead author; grey shading indicates a checklist developed for the appraisal of case series studies only.
* HTA and other reports; † checklist developed by the Review Body for Interventional Procedures, UK; ‡ used checklist published by Cho et al.[42] 1994; § information abstracted from Cho et al.[42] 1994.
Abbreviations: AACPDM: American Academy for Cerebral Palsy and Development Medicine; CASP: Critical Appraisal Skills Programme tool; CI: confidence interval; CRD: Centre for Reviews and Dissemination; K: Kappa value (interrater reliability); MSAC: Medical Services Advisory Committee; NICE: National Institute for Health and Clinical Excellence; r: interclass correlation; RCT: randomized controlled trial; W: Kendall's coefficient of concordance

Table C.2: Criteria from checklists used for the appraisal of case series studies (sorted by domain, frequency of appearance, and source)

Description of criterion	Number of checklists/studies which include the criterion	Checklists for appraisal of CS only	Checklists for appraisal of CS and studies of other designs
Study question			
Clear description of the study rationale, purpose, hypothesis, aim, objective, or any combination of these	14	**4**[10,11,13,15]	**10**[18,20-22,25,33,36,37,39,40]
Study design able to answer research question, appropriate for the aim	5	**1**[13]	**4**[18,31,39,40]
Sufficient evidence to justify the study, relevant background reviewed	3	–	**3**[22,27,40]
Specified why this case study was undertaken	1	**1**[15]	–
Type of research design employed by the study described	1	**1**[15]	–
Study population			
Eligibility criteria described (inclusion and or exclusion criteria)	23	**8**[6-11,13,15]	**15**[16-19,22-28,32,35,37,38]
Sample (adequate, described) from relevant population, appropriate methods of recruitment, size justified	17	**5**[6,7,12-14]	**12**[17,19,20,21,23,26,27,31,36,39-41]
Description of baseline characteristics; definition of participants (age, gender, etc.)	10	**2**[8,10]	**8**[18,20-22,25,35-37]
Prospective study design (data collected prospectively – before and after)	9	**5**[6,8,10-12]	**4**[14,27,28,33]
Participants recruited consecutively	8	**4**[6,10-12]	**4**[17,30,33,34]
Homogeneity/ comparability of population at entry into the study (diagnostic, prognostic, disease status or progression)	5	**1**[6]	**4**[19,23,26,37]

Selection of participants appropriate to the study question	3	–	**3**[18,27,34]
Participants entering the study at a similar point in their disease progression	3	**3**[6,7,10]	–
Recruitment period clearly stated; same time	2	**2**[6,10]	–
Case series collected in more than one centre (multicentre study)	2	**2**[10,11]	–
Adequate description of the disease/condition being treated	1	**1**[13]	–
Staff, places, facilities for treatment representative of the treatment of majority of patients	1	–	**1**[21]
Inclusion and exclusion criteria match the goals of the study	1	–	**1**[22]
Method of selection cases identified and appropriate	1	**1**[14]	–
Comorbidities described	1	–	**1**[24]
Sample size large enough to detect statistically significant differences in primary and secondary outcomes	1	–	**1**[30]
At least 50 cases included	1	**1**[8]	–
Ascertainment of exposure (secure record, structured interview)	1	–	**1**[41]
Demonstration that outcome of interest was not present at start of study	1	–	**1**[41]
Criteria (inclusion/exclusion) applied equally to all groups	1	–	**1**[37]
Study groups comparable to nonparticipants with regard to confounding factors	1	–	**1**[37]
Intervention			
Loss to follow-up (nonrespondents, withdrawals, and dropouts) reported *Note: Various ranges are reported for an acceptable loss to follow-up (5% to 25%)*	21	**4**[6,8,10,12]	**17**[16-19,21-26,33,35,36,38-41]
Description of the intervention of interest (e.g. type, setting, location(s), unit of delivery, exposure, duration)	16	**4**[8-10,13]	**12**[16,19-22,25-27,36,37,39,40]
Timing for follow-up measurements (duration of follow-up reported, appropriateness, completeness)	15	**5**[6-8,10,14]	**10**[19,24,26,27,33,34,37-39,41]
Description of co-intervention(s) received in addition to the intervention; co-interventions avoided or comparable	7	**3**[8,10,12]	**4**[19,26,31,38]

Attempt made to blind study participants to the intervention they have received	6	–	**6**[18,19,21,27,30,38]
Active follow-up; short- and long-term follow-up performed	3	–	**3**[26,29,38]
Therapy available and feasible in practice	1	–	**1**[23]
The rationale for the treatment protocol is clear	1	**1**[13]	–
Details of methods/ procedures are adequate to allow the study to be repeated	1	**1**[13]	–
Outcome measurement			
Outcomes (primary, secondary) clearly defined (timing, measurement, valid, reliable, standardized, objective criteria)	26	**8**[6-13]	**18**[16,17,19-21,25-31,35,37,38,40]
Blind assessment of outcomes; assessment independent and objective	23	**5**[6,7,10,12,13]	**18**[16-21,27-34,36-38,41]
Used objective methods to measure outcomes	4	**2**[7,10]	**2**[28,29]
Study's findings/ outcome measures respond/relevant to research objective(s)/ question(s)	3	**1**[10]	**2**[36,39]
Used subjective methods to measure outcomes	2	**1**[10]	**1**[28]
Outcome measures before and after intervention	2	**1**[10]	**1**[28]
Sufficient description of the series and the distribution of prognostic factors for comparisons of subseries	1	**1**[7]	–
Outcomes stratified (based on follow-up, etiologies, co-interventions)	1	**1**[10]	–
More than one examiner for outcome assessment; examiner calibration carried on	1	–	**1**[25]
Intervention undertaken by someone experienced in performing the procedure	1	**1**[6]	–
Interval between (different) measurements identical for all patients	1	**1**[8]	–
Exposure measured equally in all study groups	1	–	**1**[37]
Data collected are relevant and complete	1	**1**[13]	–
Type of measurements undertaken in the study mentioned	1	**1**[15]	–
All important outcomes are considered	1	**1**[15]	–

Statistical analysis			
Statistical test appropriate, valid	14	**3**[8,10,13]	**11**[16,18,19,21,22,26-28,31,37,40]
Compliance with intervention measured and noncompliers analyzed correctly	8	–	**8**[19,21-23,26,27,31,38]
Intention-to-treat analysis	8	**1**[8]	**7**[19,20,23,26,32,35,38]
Statistical significance and or clinical significance considered	6	–	**6**[22,23,25,27,36,40]
Estimates of random variability (standard error, standard deviation, confidence intervals, precision)	6	**1**[10]	**5**[20,21,26,28,39]
Contamination and co-intervention reported or avoided	6	–	**6**[22,23,26,29,31,40]
Methods and tests of statistical analysis described	4	–	**4**[18-20,25]
Size and significance of the effect reported	3	–	**3**[27,28,33]
Study has sufficient power to detect a clinically important effect where the probability value is less than 5%	3	–	**3**[21,27,28]
Point measures/ estimates	3	–	**3**[19,32,38]
Power calculation provided (P values, confidence intervals)	3	–	**3**[16,18,37]
Discussion of possible confounders	3	**1**[10]	**2**[37,39]
Actual probability values reported	2	**1**[10]	**1**[21]
Presentation of the mean of the outcome measures; percentage	2	–	**2**[26,28]
Accounted for every patient who is eligible for the study but does not enter it	2	–	**2**[22,36]
Accounted for all participants in the study	2	–	**2**[22,29]
Analysis of outcomes based on the number of patients available at the time when the follow-up measures are taken	1	**1**[10]	–
Multiple comparisons taken into consideration	1	–	**1**[37]
Modeling and multivariate techniques appropriate	1	–	**1**[37]
Results in absolute numbers when feasible	1	–	**1**[25]
Dose-response assessment if appropriate	1	–	**1**[37]
Results			
Main findings/clinically relevant outcomes clearly reported; clinical importance reported	14	**6**[6,8,10,11,13,15]	**8**[18,19,21-23,36,37,40]
Description of adverse events/side effects	11	**3**[8,10,13]	**8**[19-21,23,25,26,35,38]
Conclusions supported by objectives, analysis, and results	6	**1**[10]	**5**[18,20,27,37,39]

Generalizability	6	**1**[15]	**5**[17,20,22,28,39]
Discussion, interpretation of the results considering study hypotheses, limitations, potential biases	3	**1**[10]	**2**[20,39]
Any conclusions stated; appropriate conclusions	3	**1**[15]	**2**[25,40]
Results based on "data dredging"	2	**1**[10]	**1**[21]
Competing interest statement about type and source of support received for the study or relationship of the author(s) with the manufacturer of the technology	2	**1**[10]	**1**[37]
Other criteria			
Method of randomization, treatment allocation concealed groups similar at baseline, experimental and control interventions explicitly described, timing of outcome assessment in both groups comparable, sample size described in each group	1	–	**1**[38]

Abbreviations: CS: case series study

APPENDIX D: LIST OF DELPHI PANELISTS

The names and affiliations of the seven panelists who participated in the Delphi process are listed in alphabetical order:

Alun Cameron, PhD
Senior Research Manager
ASERNIP-S, Research & Academic Surgery Division
Australia

Paula Corabian, BA, MPH
Research Associate, Health Technology Assessment (HTA)
Institute of Health Economics
Canada

Bing Guo, MD, MSc
Research Associate, HTA
Institute of Health Economics
Canada

Christa Harstall, BScMLS, MHSA
Director, HTA
Institute of Health Economics
Canada

Iñaki Imaz Iglesia, MD, PhD, MPH
Senior Researcher
Agency for Health Technologies Assessment Institute of Health
"Carlos III", Ministry of Science and Innovation
Spain

Carmen Moga, MD, MSc
Research Associate, HTA
Institute of Health Economics
Canada

Ann Scott, BSc (Hons), PhD
Research Associate, HTA
Institute of Health Economics
Canada

APPENDIX E: RESULTS FROM DELPHI ROUNDS

Table E.1: Summary of results from Delphi rounds

Initial criterion	Decision from Delphi rounds			Final criterion (4th round)
	1st	2nd	3rd	
1. Is the hypothesis/aim/objective of the study stated in the abstract, introduction, or methods section?	✓			1. Is the hypothesis/aim/objective of the study clearly stated in the abstract, introduction or methods section?
2. Are the characteristics of the patients included in the study clearly described?	✓			2. Are the characteristics of the participants included in the study described?
3. Were the case series collected in more than one centre?		✓		3. Were the cases collected in more than one centre?
4. Are the eligibility criteria (inclusion and exclusion criteria) explicit and appropriate?	✓			4. Are the eligibility criteria (inclusion and exclusion criteria) to entry the study explicit and appropriate?
5. Were data collected prospectively?		E		–
6. Were patients recruited consecutively?	✓			5. Were participants recruited consecutively?
7. Did patients enter the study at a similar point in the disease?	✓			6. Did participants enter the study at a similar point in the disease?
8. Were the subjects recruited during the same period of time?		E		–
9. Did the authors describe the intervention?	✓			7. Was the intervention clearly described in the study?
10. In addition to intervention, did the patients receive any co-intervention?		✓		8. Were additional interventions (co-interventions) clearly reported in the study?
11. Was there loss to follow-up reported?	✓			14. Was the loss to follow-up reported?
12. Are outcomes (primary, secondary) clearly defined in the introduction or methodology section?	✓			9. Are the outcome measures clearly defined in the introduction or methods section?
13. Were the outcomes assessed blind/independent to intervention status?		E		–
14. Did the authors use accurate (standard, valid, reliable), objective methods to measure the outcomes?	✓			10. Were relevant outcomes appropriately measured with objective and/or subjective methods?
15. Did the authors use standardized subjective measures (e.g. questionnaires or patient symptoms interview forms) to assess the outcomes?		E		See criterion 10. Excluded in round 2 and reintroduced in round 3.
16. Were outcomes assessed before and after intervention?	✓			11. Were outcomes measured before and after intervention?
17. Was the length of follow-up clearly described/reported?	✓			13. Was the length of follow-up reported?
18. Were the statistical tests used to assess the primary outcomes appropriate?	✓			12. Were the statistical tests used to assess the relevant outcomes appropriate?

19. Have actual probability values been reported (e.g. p = 0.035 rather than p < 0.05) for the primary outcome measurements except where the probability value is less than 0.001?		E		–
20. Does the study provide estimates of the random variability in the data for the primary outcomes (e.g. standard error, standard deviation, confidence intervals)?	✓			15. Does the study provide estimates of the random variability in the data analysis of relevant outcomes?
21. Was there a discussion/assessment of possible confounders?		E		–
22. Was the analysis of outcomes based on the number of patients available at the time when the follow-up measures were taken? *Reformulated for the 2nd round:* Was the analysis of outcomes based on intention to treat?		✓	E	–
23. Are the main findings of the study clearly described?		E		–
24. Are outcomes of the study stratified (e.g. based on follow-up periods, etiologies, co-intervention)?		E		–
25. Do the study's findings respond to research objective(s)/question(s)?		E		–
26. Are adverse events that may be a consequence of the intervention reported?	✓			16. Are adverse events reported?
27. Are results based on data dredging?		E		-
28. Are the conclusions of the study supported by results?		✓		17. Are the conclusions of the study supported by results?
29. Are the limitations of the study taken into consideration?		E		-
30. Is there a competing interest statement about the type and source of support received for the study or about the relationship of the author(s) or other contributors with the manufacturer of the technology?		✓		18. Are both competing interest and source of support for the study reported?

Abbreviations: E: excluded. p: probability; ✓: criterion included

APPENDIX F: RESULTS – DELPHI FIRST ROUND

Table F.1: Panelists' ranking

No	Criterion	Rank†					Rank (mean)	Decision
		1	2	3	4	5		
1	**Is the hypothesis/aim/objective of the study stated in the abstract, introduction, or methods section?**	6	*1				2	**Included**
2	**Are the characteristics of the patients included in the study clearly described?**	5	*2				3	**Included**
3	Were the case series collected in more than one centre?	4	1	*1	1		7	Round 2
4	**Are the eligibility criteria (inclusion and exclusion criteria) explicit and appropriate?**	6	*1				3	**Included**
5	Were data collected prospectively?	3	3	*1			6	Round 2
6	**Were patients recruited consecutively?**	*6	1				2	**Included**
7	**Did patients enter the study at a similar point in the disease?**	*5	1		1		5	**Included**
8	Were the subjects recruited during the same period of time?	1	*4		1		10	Round 2
9	**Did the authors describe the intervention?**	5	*2				3	**Included**
10	In addition to intervention, did the patients receive any co-intervention?	*4	1	2			6	Round 2
11	**Was there loss to follow-up reported?**	*6	1				2	**Included**
12	**Are outcomes (primary, secondary) clearly defined in the introduction or methodology section?**	6	*1				2	**Included**
13	Were the outcomes assessed blind/ independent to intervention status?	2	*2		2	1	12	Round 2
14	**Did the authors use accurate (standard, valid, reliable) objective methods to measure the outcomes?**	*5	1	1			4	**Included**
15	Did the authors use standardized subjective measures (e.g. questionnaires or patient symptoms interview forms) to assess the outcomes?	2	*3	2			8	Round 2
16	**Were outcomes assessed before and after intervention?**	*6	1				2	**Included**
17	**Was the length of follow-up clearly described/reported?**	*7					1	**Included**
18	**Were the statistical tests used to assess the primary outcomes appropriate?**	*5	1	1			4	**Included**
19	Have actual probability values been reported (e.g. 0.035 rather than p< 0.05) for the primary outcome measurements except where the probability value is less than 0.001?		*4	2	1		11	Round 2
20	**Does the study provide estimates of the random variability in the data for the primary outcomes (e.g. standard error, standard deviation, confidence intervals)?**	*6		1			3	**Included**
21	Was there a discussion/assessment of possible confounders?	3	*3	1			6	Round 2
22	Was the analysis of outcomes based on the number of patients available at the time when the follow-up measures were taken?	*4	2	1			5	Round 2

23	Are the main findings of the study clearly described?	2	*4	1			7	Round 2
24	Are outcomes of the study stratified (e.g. based on follow-up periods, etiologies, co-intervention)?	2	*4		1		8	Round 2
25	Do the study's findings respond to research objective(s)/question(s)?	3	*2	2			7	Round 2
26	**Are adverse events that may be a consequence of the intervention reported?**	*6		1			3	**Included**
27	Are results based on data dredging?	1	4	*2			9	Round 2
28	Are the conclusions of the study supported by results?	4		*3			7	Round 2
29	Are the limitations of the study taken into consideration?	3	*3	1			6	Round 2
30	Is there a competing interest statement about the type and source of support received for the study or about the relationship of the author(s) or other contributors with the manufacturer of the technology?	*4	2	1			5	Round 2

* The asterisk indicates the individual rank allocated by one panelist, in this case panelist 1; † rank 1: criterion very important; rank 2: criterion somewhat important; rank 3: criterion equivocal; rank 4: criterion not very important; rank 5: criterion not important at all; ‡ higher standard deviations show greater levels of disagreement among panelists on the criterion.

Abbreviations: SD: standard deviation

Results shaded in grey indicate at least 70% agreement among panelists (at least 5 out of 7 panelists judged the criterion very important (rank 1) or not important (rank 5)). This represents the cut-off for including or excluding criteria.

Bolded statements refer to criteria included after the first round.

Table F.2: Panelists' responses (means, standard deviations, and correlation with the mean vector, and their rankings)

Panelist	Mean (Rank)	SD‡ (Rank)	Correlations (Rank)
1	1.77 (6)	0.73 (3)	0.43 (5)
2	1.40 (2)	0.89 (5)	0.48 (4)
3	2.30 (7)	1.15 (7)	0.68 (2)
4	1.47 (4)	0.51 (2)	0.66 (3)
5	1.57 (5)	0.82 (4)	0.41 (6)
6	1.45 (3)	0.91 (6)	0.72 (1)
7	1.37 (1)	0.49 (1)	0.34 (7)

‡ Larger standard deviations show a wider use of the full scale (1 to 5) by a panelist
Abbreviations: SD: standard deviation

Table F.3: Suggestions made by panelists

Suggestions
Add: "Was each intervention carried out by a named surgeon(s)?"
Add: "Were there similarities between enrolled participants and those not enrolled?"
Add: "Did the study mention power of calculation?"
Criteria 24 and 27 are duplicates. Any stratification analysis should be predefined; otherwise, it's data dredging. Thus, 24 and 27 should be combined; e.g. "Were subgroup analyses defined a priori?"
Criteria 25 and 28 appeared to be repeated as these criteria assess similar questions.
Was the length of follow-up appropriate? E.g. short follow-up periods may not be appropriate for interventions in patients with chronic disease.
For criterion 8, an additional piece of information is how long that time span is. The patients may have been recruited over the same time period, but if that time period spanned 20 years, then there's the issue that earlier patients probably received different management, and even treatment regimens, compared to later patients. Perhaps an additional question would be "Was the time span appropriate (short enough to rule out significant practice variation that may be a confounder)?"
Criterion 22 should include terms like "intention to treat analyses".
No suggestion for new criteria because these 30 criteria include all questions ever worked.
Should make it clearer that this tool is for intervention studies only, not diagnostic or other study types.
The checklist should adjunct explanations for each criterion.

APPENDIX G: RESULTS – DELPHI SECOND ROUND

Table G.1: Panelists' ranking

No	Criterion	Rank†					Rank (mean)	Decision
		1	2	3	4	5		
3	**Were the case series collected in more than one centre?**	5		*1	1		5	**Included**
5	Were data collected prospectively?	4	*3				3	Excluded
8	Were the subjects recruited during the same period of time?	*4	1		2		7	Excluded
10	**In addition to intervention, did the patients receive any co-interventions?**	*6		1			2	**Included**
13	Were the outcomes assessed blind/ independent to intervention status?	1	*2		3	1	10	Excluded
15	Did the authors use standardized subjective measures (e.g. questionnaires or patient symptoms interview forms) to assess the outcomes?	2	*3	1		1	9	Excluded
19	Have actual probability values been reported (e.g. p = 0.035 rather than p < 0.05) for the primary outcome measurements except where the probability value is less than 0.001?		*6	1			8	Excluded
21	Was there a discussion/assessment of possible confounders?	3	*4				4	Excluded
22	**Was the analysis of outcomes based on the number of patients available at the time when the follow-up measures were taken?**	*5	2				3	**Included**
23	Are the main findings of the study clearly described?	1	3	*3			9	Excluded
24	Are outcomes of the study stratified (e.g. based on follow-up periods, etiologies, co-intervention)?	3	*4				4	Excluded
25	Do the study's findings respond to research objective(s)/question(s)?	3	*2	2			6	Excluded
27	Are results based on data dredging?	2	2	*2	1		9	Excluded
28	**Are the conclusions of the study supported by results?**	5		*2			4	**Included**
29	Are the limitations of the study taken into consideration?	3	*4				4	Excluded

30	**Is there a competing interest statement the type and source of support received for the study or the relationship of the author(s) or other contributors with the manufacturer of the technology?**	*6	1				1	**Included**

* The asterisk indicates the individual rank allocated by one panelist, in this case panelist 1; † rank 1: criterion very important; rank 2: criterion somewhat important; rank 3: criterion equivocal; rank 4: criterion not very important; rank 5: criterion not important at all; ‡ higher standard deviations show greater levels of disagreement among panelists on the criterion.

Abbreviations: SD: standard deviation

Results shaded in grey indicate at least 70% agreement among panelists (at least 5 out of 7 panelists judged the criterion very important (rank 1) or not important (rank 5)). This represents the cut-off for including or excluding criteria.

Bolded statements refer to criteria included after the second round.

Table G.2: Panelists' responses (means, standard deviations, and correlation with the mean vector, and their rankings)

Panelist	Mean (Rank)	SD (Rank)	Correlations (Rank)
1	2.06 (6)	0.68 (2)	0.31 (6)
2	1.69 (3)	1.14 (6)	0.56 (4)
3	2.50 (7)	1.03 (5)	0.66 (3)
4	1.88 (5)	0.96 (4)	0.68 (2)
5	1.50 (2)	0.73 (3)	0.50 (5)
6	1.75 (4)	1.34 (7)	0.69 (1)
7	1.44 (1)	0.51 (1)	0.12 (7)

‡ Larger standard deviations show a wider use of the full scale (1 to 5) by a panelist
Abbreviations: SD: standard deviation

Table G.3: Results and decision for suggestions made by panelists in the first round

Suggested new criteria	Yes	No	NA	Comments	Decision
Was each intervention carried out by a named surgeon(s)?	0	5	2	- Unclear – same "surgeon" cannot be at several centres. - This criterion is not very clear.	Rejected
Were there similarities between enrolled participants and those not enrolled?	2	5	0	- This works for case/control studies - It is difficult to respond to this question based on information from a case series study. It is important that the publication mention the inclusion/exclusion criteria.	Rejected
Did the study mention power of calculation?	3	4	0	- I think that the checklist includes questions on statistics.	Rejected
Suggested refinement of existed criteria					
Criteria 24 and 27 to be combined as follows: "Were subgroup analyses defined a priori?"	6	0	1	- Both 24 and 27 criteria are important. I agree with the suggestion to combine them. - A proposal: Were subgroup analyses appropriate and defined a priori?	Rejected (Note: Both criteria were excluded; see Table E1)
Criteria 25 and 28 are similar. Any suggestions on how to combine the two?	3	3	1	- Criterion 25 may also be covered by the a priori question above. Question 28 is more about conclusions – the results themselves may be of more importance. - There are two separate issues, and they cannot be combined. Criterion 25 concerns whether the study measured outcomes that were in line with the stated objectives of the study. This helps identify studies that measure only those things that have positive effects; e.g. if a study's objective was to determine efficacy but they only measure a change in blood levels of something rather than a clinically useful measure such as function or quality of life. In contrast, criterion 28 is asking whether the results match the conclusions. This criterion is designed to weed out studies where, for example, the authors have made very positive conclusions based on equivocal results. This is not the same issue as what criterion 25 is dealing with, which is why they should not be combined. - Include only criterion 28.	Rejected
Was the length of follow-up appropriate?	4	3	0	- Combine with criterion 17. - It is difficult to define the appropriate follow-up period. The length of follow-up should be specific/different for different interventions.	Rejected

Criterion 8 to be changed to "Were the subjects recruited during the same period of time? How long is that time span? Was the time span appropriate?"	2	3	2	- Too many questions for one criterion. - A proposal: Were the subjects recruited during an appropriate and similar time span? Include an explanation.	Rejected
Criterion 22 must include terms like "intention to treat analyses". Any suggestion on how to reword this criterion?	2	5	0	- This may be more appropriate for a controlled trial than for a case series. - Did the study include an intention-to-treat analysis? - Was the analysis of outcomes based on "intention to treat"? - Was the analysis performed by intention to treat?	Rejected
Suggested instruction on how to use the checklist					
This tool should be used for intervention studies only, not for diagnostic or other study types.	7	0	0	- Or at least make it clear that this list would need to be modified for use with other study types.	**Accepted**
The checklist must adjunct explanations for each criterion.	6	1	0	- I presume this means that the list should be operationalized with instructions – this is the next step after sorting out the criteria anyway. - The list should be customized by reviewers, depending on the subject.	**Accepted**

Abbreviations: NA: no response

APPENDIX H: RESULTS – DELPHI THIRD ROUND

Table H.1: Quality appraisal criteria selected from the first two rounds

Crt. No.	Criterion included	Number of panelists indicating the criterion can be eliminated
1.	Is the hypothesis/aim/objective of the study stated in the abstract, introduction, or methods section?	–
2.	Are the characteristics of the patients included in the study clearly described?	–
3.	Were the case series collected in more than one centre?	2
4.	Are the eligibility criteria (inclusion and exclusion criteria) explicit and appropriate?	–
5.	Were patients recruited consecutively?	–
6.	Did patients enter the study at a similar point in the disease?	1
7.	Did the authors describe the intervention?	–
8.	In addition to intervention, did the patients receive any co-interventions?	3
9.	Was loss to follow-up reported?	–
10.	Are outcomes (primary, secondary) clearly defined in the introduction or methodology section?	2
11.	Did the authors use accurate (standard, valid, reliable) objective methods to measure the outcomes?	1
12.	Were outcomes assessed before and after intervention?	–
13.	Was the length of follow-up clearly described/reported?	–
14.	Were the statistical tests used to assess the primary outcomes appropriate?	–
15.	Does the study provide estimates of the random variability in the data for the primary outcomes (e.g. standard error, standard deviation, confidence intervals)?	–
16.	Was the analysis of outcomes based on intention to treat?	**5 (71%)***
17.	Are adverse events that may be a consequence of the intervention reported?	–
18.	Are the conclusions of the study supported by results?	2
19.	Is there a competing interest statement about the type and source of support received for the study or about the relationship of the author(s) or other contributors with the manufacturer of the technology?	–

* Indicates at least 70% agreement among panelists; at least 5 out of 7 panelists judged the criterion very important (rank 1) or not important (rank 5). This represents the cut-off for including or excluding criteria.

Table H.2: Quality appraisal criteria excluded from the first two rounds

Criterion excluded	Number of panelists indicating the criterion should be reconsidered
1. Were data collected prospectively?	2
2. Were the subjects recruited during the same period of time?	–
3. Were the outcomes assessed blind/independent to intervention status?	–
4. Did the authors use standardized subjective measures (e.g. questionnaires or patient symptoms interview forms) to assess the outcomes?	2
5. Have actual probability values been reported (e.g. p = 0.035 rather than p < 0.05) for the primary outcome measurements except where the probability value is less than 0.001?	–
6. Was there a discussion/assessment of possible confounders?	–
7. Are the main findings of the study described clearly?	–
8. Are outcomes of the study stratified (e.g. based on follow-up periods, etiologies, co-intervention)?	–
9. Do the study's findings respond to research objective(s)/question(s)?	–
10. Are results based on data dredging?	–
11. Are the limitations of the study taken into consideration?	–

APPENDIX I: CRITERIA AND DRAFT DICTIONARY FOR THE QUALITY ASSESSMENT CHECKLIST

Study objective

1. Is the hypothesis/aim/objective of the study clearly stated in the abstract, introduction, or methods section?

Yes: The hypothesis/aim/objective of the study is clearly stated in the abstract, introduction, or methods section.
No: The hypothesis/aim/objective is not provided in the abstract, introduction, or methods section.

Study population

2. Are the characteristics of the participants included in the study described?

Yes: The most relevant characteristics are presented. The authors should report the total number, age, and gender distribution of the participants. Ethnicity, severity of disease/condition, comorbidity, or etiology should also be included, if relevant.
No: The most relevant characteristics of the participants are not reported. If only the number of participants was reported or any of the relevant characteristics is missing, the question should be answered no.
Note: Assessor(s) should decide which aspects are important before using the checklist.

3. Were the cases collected in more than one centre?

Yes: Cases are collected in more than one centre (multicentre study).
No: Cases are collected from one centre, or it is unclear where patients came from.

4. Are the eligibility criteria (inclusion and exclusion criteria) to entry the study explicit and appropriate?

Yes: The eligibility criteria are clearly stated and replicable, and match the objective of the study.
No: The eligibility criteria are not clearly stated.
Note: Assessor(s) should decide which aspects are important before using the checklist.

5. Were participants recruited consecutively?

Yes: There is a clear statement that the participants are recruited consecutively.
No: The participants were recruited based on other criteria, such as access to intervention determined by the distance or availability of resources. The method used to recruit participants is not clearly stated.

6. Did participants enter the study at a similar point in the disease?

Yes: There is a clear description about the clinical status of participants, duration of condition (exposure) before the intervention, comorbidity, severity, or complications of all participants in the study.
No: There is no description about whether participants entered the study at a similar point in the disease. Participants did not enter the study at a similar point in the disease, as revealed by a wide range of disease duration before entering the study or different comorbidities or complications due to progression of their condition/disease.
Note: Assessor(s) should decide which aspects are important before using the checklist.

Intervention and co-intervention

7. Was the intervention clearly described in the study?

Yes: There is a detailed description about the characteristics of the intervention (e.g. dosage, frequency of administration, duration, permanent or temporary intervention, and technical parameters/characteristics of a device).
No: The intervention is only mentioned by name without any details, the information provided is unclear, or important parameters of the intervention are missing from the presentation.
Note: Assessor(s) should decide which aspects are important before using the checklist.

8. Were additional interventions (co-interventions) clearly reported in the study?

Yes: The name or type of co-intervention is acknowledged in the study. The question should be answered yes if it is obvious (based on study context) that co-interventions were unnecessary.

No: Co-intervention(s) are not reported, or the name(s) or type(s) of co-intervention(s) are unclear.
Note: Assessor(s) should decide which aspects are important before using the checklist.

Outcome measures

9. Are the outcome measures clearly defined in the introduction or methods section?

Yes: All relevant (primary and secondary) outcomes that match the objective(s) of the study are described in the introduction or methods section (e.g. accomplished, measurable improvements or effects, symptoms relieved, improved function, improved test scores, and quality of life measures).
No: The outcomes are reported for the first time in the results or conclusion section of the study.
The relevant outcomes are briefly mentioned without any details in the results, discussion, or conclusion section(s).
The outcomes reported are not relevant to study objective(s).
Note: Assessor(s) should decide which aspects are important before using the checklist.

10. Were relevant outcomes appropriately measured with objective and/or subjective methods?

Yes: Appropriate methods used to measure the outcomes are described in the methods section. These measures might be objective (e.g. gold standard tests or standardized clinical tests), and/or subjective (e.g. self-administered questionnaires, standardized forms, or patient symptoms interview forms).
No: No details are provided on the objective or subjective methods used to measure study's outcomes.

11. Were outcomes measured before and after intervention?

Yes: The relevant outcomes are measured before and after applying the intervention.
No: The outcomes are measured only after applying the intervention.

Statistical analysis

12. Were the statistical tests used to assess the relevant outcomes appropriate?

Yes: The statistical tests are clearly described in the methods section and are used appropriately (e.g. parametric test for normally distributed population vs. nonparametric test for non-Gaussian population).
No: The statistical tests used to assess the relevant outcomes are inappropriate. From the information available it is unclear the distribution of the population from which the participants at the study were selected.

Results and conclusions

13. Was the length of follow-up reported?

Yes: The length of follow-up is clearly reported.
No: The length of follow-up is not reported, or the duration of the study is unclear.

14. Was the loss to follow-up reported?

Yes: The number or proportion of patients lost to follow-up is reported.
No: The number or proportion of patients lost to follow-up is not reported.

15. Does the study provide estimates of the random variability in the data analysis of relevant outcomes?

Yes: The study reports estimates of the random variability (e.g. standard error, standard deviation, confidence intervals) for all relevant primary and secondary outcomes.
No: Estimates of the random variability are not reported for all relevant outcomes. The presentation of the random variability is unclear (e.g. measure of dispersion reported without indicating if it is standard deviation or standard error).

16. Are adverse events reported?

Yes: The undesirable or unwanted consequences of the intervention during the study period or within a prespecified time period are reported. Absence of any adverse event(s) is acknowledged in the study.
No: There is no statement about the presence or absence of adverse events.

17. Are the conclusions of the study supported by results?

Yes: The main conclusions of the study are supported by the evidence presented in the results section.

No: The conclusions are not supported by the evidence presented in the results section.

Competing interest and source of support

18. Are both competing interest and source of support for the study reported?

Yes: Both competing interest and source of support (financial or other) received for the study are reported, or the absence of any competing interest and source of support is acknowledged.
No: Either there is no information available about competing interests and sources of support, or only one of these elements is reported.

APPENDIX J: ADAPTATION OF THE DRAFT DICTIONARY

Table J.1: Adaptation of the draft dictionary

Criterion and draft dictionary	Specific aspects added for clarification
2. Are the characteristics of the participants included in the study described? **Yes**: The most relevant characteristics are presented. The authors should report the total number, age, and gender distribution of the participants. Ethnicity, severity of disease/condition, comorbidity, or etiology should also be included, if relevant. **No**: The most relevant characteristics of the participants are not reported. If only the number of participants was reported or any of the relevant characteristics is missing, the question should be answered no. *Note: Assessor(s) should decide which aspects are important before using the checklist.*	To score yes, the total number of patients, age, gender, duration of diabetes, and renal function should be mentioned. Score yes if relevant data were reported in previously published articles and these data are available. *No further discussion*
4. Are the eligibility criteria (inclusion and exclusion criteria) to entry the study explicit and appropriate? **Yes**: The eligibility criteria are clearly stated and replicable, and match the objective of the study. **No**: The eligibility criteria are not clearly stated. *Note: Assessor(s) should decide which aspects are important before using the checklist.*	To score yes, age, duration of diabetes, severe hypoglycemia or hypoglycemia unawareness, no kidney disease, and no previous kidney transplantation should be mentioned. Score yes if relevant data were reported in previously published articles and these data are available. *No further discussion*
6. Did participants enter the study at a similar point in the disease? **Yes**: There is a clear description about the clinical status of participants, duration of condition (exposure) before the intervention, comorbidity, severity, or complications of all participants in the study. **No**: There is no description about whether participants entered the study at a similar point in the disease. Participants did not enter the study at a similar point in the disease, as revealed by a wide range of disease duration before entering the study or different comorbidities or complications due to progression of their condition/disease. *Note: Assessor(s) should decide which aspects are important before using the checklist.*	To score yes, all of the following criteria should be met: (1) all patients have diabetes ≥ 5 years, (2) ≥ 80% of patients have severe unawareness, (3) ≥ 80% of patients have no kidney disease. Score yes if relevant data were reported in previously published articles and these data are available. The question should be scored no if (1) not all criteria above are met, (2) no clear description of these characteristics is provided, (3) information is not available for one of these characteristics, or (4) there is no statement that patients' characteristics were described earlier. *Discussion: Two reviewers found it difficult to determine which aspects are most important in terms of the similarity in the course of disease. There is no commonly used classification system for diabetes in terms of the severity or stage. Modification of the dictionary for this criterion was first made based on the background information that provides understanding of the most important clinical characteristics of the target population. Three aspects were considered important in terms of similarity in the course of disease: duration of diabetes; presence of hypoglycemia, hypoglycemia unawareness, or both; and presence of kidney disease. However, two researchers still found it difficult to rate this criterion, even with the modified dictionary.*

	A second modification was made to quantify the three clinical aspects: *(1) All patients had diabetes ≥ 5 years; this period was chosen because the hormone interregulatory system (i.e. glucagons and autonomic nervous system) can be impaired, and microalbuminuria can occur 5 years after diagnosis of type 1 diabetes;* *(2) ≥ 80% of the patients had severe hypoglycemia or hypoglycemia unawareness;* *(3) ≥ 80% of patients did not have kidney disease.*
7. Was the intervention clearly described in the study? **Yes**: There is a detailed description about the characteristics of the intervention (e.g. dosage, frequency of administration, duration, permanent or temporary intervention, and technical parameters/characteristics of a device). **No**: The intervention is only mentioned by name without any details, the information provided is unclear, or important parameters of the intervention are missing from the presentation. *Note: Assessor(s) should decide which aspects are important before using the checklist*	To score yes, the number of islet/per infusion, frequency of perfusion, and immunosuppressive therapy (drug, dosage, and frequency of administration) should be mentioned. Score yes if relevant data were reported in previously published articles and these data are available. *No further discussion*
8. Were additional interventions (co-interventions) clearly reported in the study? **Yes**: The name or type of any co-intervention is acknowledged in the study. The question should be answered yes if it is obvious (based on study context) that co-interventions were unnecessary. **No**: Co-intervention(s) are not reported, or name(s) or type(s) .of co-intervention(s) are unclear. *Note: Assessor(s) should decide which aspects are important before using the checklist.*	*Co-intervention* was defined as intervention(s) other than islet transplantation and immunosuppression therapy given to patients during the study; it may include diet change, exercise, insulin therapy, or antiviral and antimicrobial prophylaxis therapy to prevent/control adverse events. Score yes if relevant data were reported in previously published articles and these data are available. *No further discussion*
9. Are the outcome measures clearly defined in the introduction or methodology section? **Yes**: All relevant (primary and secondary) outcomes that match the objective(s) of the study are described in the introduction or method section (e.g. accomplished, measurable improvements or effects, symptoms relieved, improved function, improved test scores, and quality of life measures). **No**: The outcomes are reported for the first time in the results or conclusion section of the study. The relevant outcomes are mentioned briefly without any details in the results, discussion, or conclusion section(s). The outcomes reported are not relevant to study objective(s). *Note: Assessor(s) should decide which aspects are important before using the checklist*	Relevant outcomes include insulin independence or reduction of insulin, c-peptide secretion, HbA1C level (i.e. glicosylated hemoglobin), hypoglycemia episode, secondary complications, or quality of life. *No further discussion*
12. Were the statistical tests used to assess the relevant outcomes appropriate? **Yes**: The statistical tests are clearly described in the methods section and are used appropriately (e.g. parametric test for normally distributed population vs. nonparametric test for non-Gaussian population). **No**: The statistical tests used to assess the relevant	To score yes, the method used for statistical analysis should be clearly described, or the study should have stated that the statistical analysis was not performed for specific reasons such as small sample size. *Discussion: The three researchers recognized a need for a thorough group discussion about this*

outcomes are inappropriate. From the information available the distribution of the population from which the study participants were selected is unclear.	*criterion; however, time constraints did not permit this. It was then decided that a study should be scored yes if it clearly described the method used for statistical analysis or stated that the statistical analysis was not performed for a specific reasons (e.g. small sample size). In other words, the modified definition did not directly address the appropriateness of statistical analyses.*

APPENDIX K: SUGGESTIONS FOR THE CHECKLIST AND DRAFT DICTIONARY

Table K.1: Suggested modifications to the checklist and draft dictionary

Checklist and draft dictionary (Delphi process)	Modifications
1. Is the hypothesis/aim/objective of the study clearly stated in the abstract, introduction or methods section? **Yes**: The hypothesis/aim/objective of the study is clearly stated in the abstract, introduction, or methods section. **No**: The hypothesis/aim/objective is not provided in the abstract, introduction, or methods section.	**Is the hypothesis/aim/objective of the study clearly stated?** **Yes**: The hypothesis/aim/objective of the study is clearly reported. **Unclear**: The hypothesis/aim/objective of the study is vague or unclearly reported. **No**: The hypothesis/aim/objective is not reported.
2. Are the characteristics of the participants included in the study described? **Yes**: The most relevant characteristics are presented. The authors should report the total number, age, and gender distribution of the participants. Ethnicity, severity of disease/condition, comorbidity, or etiology should also be included, if relevant. **No**: The most relevant characteristics of the participants are not reported. If only the number of participants was reported or any of the relevant characteristics is missing, the question should be answered no. *Note: Assessor(s) should decide which aspects are important before using the checklist.*	**Are the characteristics of the participants included in the study described?** **Yes**: The most relevant characteristics of the participants are reported (e.g. the total number, age, and gender distribution). Ethnicity, severity of disease/condition, comorbidity, or etiology should also be included, if relevant. **Partially reported**: Only the number of participants was reported. **No**: None of the relevant characteristics of the participants is reported. *Note: Assessor(s) should decide which aspects are important before using the checklist.*
3. Were the cases collected in more than one centre? **Yes**: Cases are collected in more than one centre (multicentre study). **No**: Cases are collected from one centre or it is unclear where patients came from.	**Were the cases collected in more than one centre?** **Yes**: Cases are collected in more than one centre (multicentre study). **Unclear**: Unclear where the patients come from (i.e. single or multicentre study). **No**: Cases are collected from one centre.
4. Are the eligibility criteria (inclusion and exclusion criteria) to entry the study explicit and appropriate? **Yes**: The eligibility criteria are clearly stated and replicable, and match the objective of the study. **No**: The eligibility criteria are not clearly stated. *Note: Assessor(s) should decide which aspects are important before using the checklist.*	**Are the eligibility criteria (i.e. inclusion and exclusion criteria) for entry into the study clearly stated?** **Yes**: Both inclusion and exclusion criteria are reported. **Partially reported**: Only one, the inclusion or exclusion criteria is reported. **No**: Neither inclusion nor exclusion criteria are reported. *Note*: *Assessor(s) should decide which aspects are important before using the checklist.*
5. Were participants recruited consecutively? **Yes**: There is a clear statement that the participants are recruited consecutively. **No**: The participants were recruited based on other criteria such as access to intervention determined by the distance or availability of resources. The method used to recruit participants is not clearly stated.	**Were participants recruited consecutively?** **Yes**: There is a clear statement or it is clear from the context that the participants were recruited consecutively or study stated that all eligible patients were recruited. **Unclear**: The method used to recruit participants is not clearly stated or no information is provided about the method used to recruit participants in the study. **No**: The cases studied were a subgroup of those treated with no evidence to show that they were selected consecutively. The participants were recruited based on other criteria such as access to intervention determined by the distance or availability of resources.

6. Did participants enter the study at a similar point in the disease? **Yes**: There is a clear description about the clinical status of participants, duration of condition (exposure) before the intervention, co-morbidity, severity, or complications of all participants in the study. **No**: There is no description about whether participants entered the study at a similar point in the disease. Participants did not enter the study at similar point in the disease, as revealed by a wide range of disease duration before entering the study or different co-morbidities or complications due to progression of their condition/disease. *Note: Assessor(s) should decide which aspects are important before using the checklist.*	**Did participants enter the study at a similar point in the disease?** **Yes**: There is a clear description about all participants entering the study at a similar point in the condition/ disease based on their clinical status, duration of condition or exposure before the intervention, severity of disease, and presence of co-morbidities or complications. **Unclear**: There is no description of the characteristics of participants before entering the study or there is no statement about entering the study at a similar point in the disease. **No**: Participants did not enter the study at a similar point in the condition/disease. This can be revealed by a wide range of disease durations before entering the study or different levels of severities or comorbidities or complications due to progression of their condition/disease. *Note: Assessor(s) should decide which aspects are important before using the checklist. It might be useful to discuss with specialists to determine the most important aspects that should be considered.*
7. Was the intervention clearly described in the study? **Yes**: There is a detailed description about the characteristics of the intervention (e.g. dosage, frequency of administration, duration, permanent or temporary intervention, and technical parameters/characteristics of a device). **No**: Intervention is only mentioned by name without any details; or the information provided is unclear; or important parameters of the intervention are missing from the presentation. *Note: Assessor(s) should decide which aspects are important before using the checklist.*	**Was the intervention of interest clearly described?** **Yes**: The most relevant characteristics of the intervention are reported (e.g. dosage, frequency of administration, duration, permanent or temporary intervention, technical parameters/ characteristics of a device). **Partially reported**: Intervention is only mentioned by name. **No**: None of the relevant characteristics of the intervention was reported. *Note: Assessor(s) should decide which aspects are important before using the checklist.*
8. Were additional interventions (co-interventions) clearly reported in the study? **Yes**: The name or type of any co-intervention is acknowledged in the study. The question should be answered yes if it is obvious (based on study context) that co-interventions were unnecessary. **No**: Co-intervention(s) are not reported, or the name(s) or type(s) of co-intervention(s) are unclear. *Note: Assessor(s) should decide which aspects are important before using the checklist.*	**Were additional interventions (co-interventions) reported in the study?** **Yes**: Participants received additional co-intervention(s). **Unclear**: It is suspected that a co-intervention was administered but the information is not reported. **No**: There is a clear statement or it is clear from the context that a co-intervention was not administered. *Note: Assessor(s) should decide which aspects are important before using the checklist.*
9. Are the outcome measures clearly defined in the introduction or methods section? **Yes**: All relevant (primary and secondary) outcomes that match the objective(s) of the study are described in the introduction or methods section (e.g. accomplished, measurable improvements or effects, symptoms relieved, improved function, improved test scores, and quality of life measures). **No**: The outcomes are reported for the first time in the results or conclusion section of the study. The relevant outcomes are briefly mentioned without any details in the results, discussion, and/or conclusion section(s).	**Are the outcome measures established a priori?** **Yes**: All relevant outcome measures are reported in the introduction or methods section (e.g. accomplished, measurable improvements or effects, symptoms relieved, improved function, improved test scores, and quality of life measures). **Partially reported:** Some of the relevant outcomes are briefly reported in the introduction or methods section. **No**: The outcome measures are reported for the first time in the results, discussion, or conclusion section of the study. *Note: Assessor(s) should decide which aspects are*

The outcomes reported are not relevant to study objective(s). *Note: Assessor(s) should decide which aspects are important before using the checklist.*	*important before using the checklist.*
10. Were relevant outcomes appropriately measured with objective and/or subjective methods? **Yes**: Appropriate methods used to measure the outcomes are described in the methods section. These measures might be objective (e.g. gold standard tests or standardized clinical tests), and/or subjective (e.g. self-administered questionnaires, standardized forms, or patient symptoms interview forms). **No**: No details are provided on the objective and/or subjective methods used to measure study's outcomes.	**Were the relevant outcomes measured with appropriate objective and/or subjective methods?** **Yes**: All relevant outcomes are measured with appropriate methods which are described in the methods section. These measures might be objective (e.g. gold standard tests or standardized clinical tests), subjective (e.g. self-administered questionnaires, standardized forms, or patient symptoms interview forms), or both. **Unclear**: It is unclear how the relevant outcomes were measured. No information is provided on the methods used to measure study's relevant outcomes. **No**: The methods used to measure outcomes were inappropriate. *Note: Assessor(s) should decide which methods are appropriate before using the checklist. Appropriate methods are defined as commonly used assays/methods to measure the outcome of interest.*
11. Were outcomes measured before and after intervention? **Yes**: The relevant outcomes are measured before and after applying the intervention. **No**: The outcomes are measured only after applying the intervention.	**Were the relevant outcomes measured before and after the intervention?** **Yes**: The relevant outcomes are measured before and after applying the intervention. **Unclear**: It is unclear when the outcomes were measured. **No**: The study reported only outcomes measured after applying the intervention.
12. Were the statistical tests used to assess the relevant outcomes appropriate? **Yes**: The statistical tests are clearly described in the methods section and are used appropriately (e.g. parametric test for normally distributed population vs. nonparametric test for non-Gaussian population). **No**: The statistical tests used to assess the relevant outcomes are inappropriate. From the information available it is unclear the distribution of the population from which the participants at the study were selected.	**Were the statistical tests used to assess the relevant outcomes appropriate?** **Yes**: The statistical tests are clearly described in the methods section of the study and are used appropriately (e.g. parametric test for normally distributed population vs. nonparametric test for non-Gaussian population). The reviewer should assign a yes score if no statistical analysis was performed but reasons for this were stated. **Unclear**: The statistical tests are not described in the methods section of the study or there is no information about the statistical analysis. **No**: The statistical tests were used inappropriately. *Note: Assessor(s) should decide which statistical tests are appropriate before using the checklist. Request expert(s) assistance if necessary.*
13. Was the length of follow-up reported? **Yes**: The length of follow-up is clearly reported. **No**: The length of follow-up is not reported, or duration of the study is unclear.	**Was the length of follow-up reported?** **Yes**: The length of follow-up is clearly reported (mean, median, range, standard deviation). **Unclear**: The duration of follow-up is not clearly reported. **No**: The length of follow-up is not reported.
14. Was the loss to follow-up reported? **Yes**: The number or proportion of patients lost to follow-up is reported. **No**: The number or proportion of patients lost to follow-up	**Was the loss to follow-up reported?** **Yes**: The number or proportion of participants lost to follow-up is clearly reported or authors report outcome results on all participants included initially, or number

is not reported.	lost to follow-up can be subtracted from the number enrolled and number analyzed. **Unclear**: It is not clear from the information provided how many participants were lost to follow-up or it is an inconsistence of reporting lost to follow-up (e.g. discrepancies between information from tables and text). **No**: The number or proportion of participants lost to follow-up is not reported.
15. Does the study provide estimates of the random variability in the data analysis of relevant outcomes? **Yes**: The study reports estimates of the random variability (e.g.; standard error, standard deviation, confidence intervals) for all relevant primary and secondary outcomes. **No**: Estimates of the random variability are not reported for all relevant outcomes. The presentation of the random variability is unclear (e.g. measure of dispersion reported without indicating if it is standard deviation or standard error).	**Does the study provide estimates of the random variability in the data analysis of relevant outcomes?** **Yes**: The study reports estimates of the random variability (e.g. standard error, standard deviation, confidence interval for parametric data, and range and interquartile range for nonparametric data) for all relevant outcomes. **Unclear or partially reported**: The presentation of the random variability is unclear (e.g.; the measure of dispersion is reported without indicating if it is a standard deviation or standard error). Estimates of the random variability are not reported for all relevant outcomes. No: The study does not report estimates of the random variability.
16. Are adverse events reported? **Yes**: The undesirable or unwanted consequences of the intervention during the study period or within a prespecified time period are reported. The absence of adverse event(s) is acknowledged in the study. **No**: There is no statement about the presence or absence of adverse events.	**Are the adverse events related with the intervention reported?** **Yes**: The undesirable or unwanted consequences of the intervention during the study period or within a pre-specified time period are reported. The absence of adverse event(s) is acknowledged in the study. **Partially reported**: It is deducible that only some but not all potential adverse events are reported. **No**: There is no statement about the presence or absence of adverse events. *Note: Assessor(s) should decide what are the most important adverse events before using the checklist*.
17. Are the conclusions of the study supported by results? **Yes**: The main conclusions of the study are supported by the evidence presented in the results section. **No**: The conclusions are not supported by the evidence presented in the results section.	**Are the conclusions of the study supported by results?** **Yes**: The conclusions of the study (in terms of patient, intervention, outcomes) are supported by the evidence presented in the results and discussion sections. **Partially reported**: Not all components of the patient, intervention, outcomes are supported by the evidence presented in the results and discussion section. **No**: The conclusions are not supported by the evidence presented in the results and discussion section.
18. Are both competing interest and source of support for the study reported? **Yes**: Both competing interest and source of support (financial or other) received for the study are reported or the absence of any competing interest and source of support is acknowledged. **No**: Either there is no information available about competing interest and source of support or only one of these elements is reported.	**Are both competing interests and sources of support for the study reported?** **Yes**: Both competing interests and sources of support (financial or other) received for the study are reported, or the absence of any competing interest and source of support is acknowledged. **Partially reported**: Only one of these elements is reported. **No**: Neither competing interests nor sources of support was reported.

	New proposed criterion: Was the study conducted prospectively? **Yes:** It is clearly stated that the study was conducted prospectively. **Unclear**: The design of the study is not mentioned or it is unclear if the study was conducted prospectively. **No:** The authors clearly stated that it was a retrospective study.
	New proposed criterion: Were the relevant outcomes assessed blinded to intervention status? **Yes:** The relevant outcomes were analyzed by individuals who were not aware of the intervention status. **Unclear**: The study did not report whether the outcome assessors were aware of the intervention status. **No:** It is clearly stated or obvious that the relevant outcomes were analyzed by individuals who were aware of the intervention status.

Note: Assessor(s) may decide to use a cut-off point to separate studies into high and low quality based on the number of criteria from the checklist met. Alternatively, they might identify some criteria from the checklist which are most relevant to a specific project and can focus more on discussing the outcomes of studies that meet those selected criteria.

A criterion should receive a yes score if information was reported in another publication and the reviewer retrieved and checked that publication.

APPENDIX L: PRELIMINARY FACTOR ANALYSIS

Reduction of the criteria by study matrix and graphical presentation of second and third factors of variation

The matrix containing the average ratings for each of the 18 criteria for each of the 13 studies rated was subjected to a singular value decomposition (analysis done with an adapted program to further describe the pattern of differences across criteria and across studies).[99] Figure L.1 shows the major dimension along which variation occurred. The studies (labelled s1 to s13) are ordered from the highest average rating (across the 18 criteria) to the lowest. Similarly, the criteria/items (i1 to i18) are ordered from highest (across the 13 studies) to lowest. For analytical reasons, these orderings do not correspond exactly to the ordering presented in Tables 7 and 8, though the differences are slight. Figure L.1 shows the relative difference between these units; a higher value of the column indicates that the study or item scored better in comparison to the other similar units.

Figure L.1: First dimension of study by criterion/item matrix (SigmaPlot® 11)

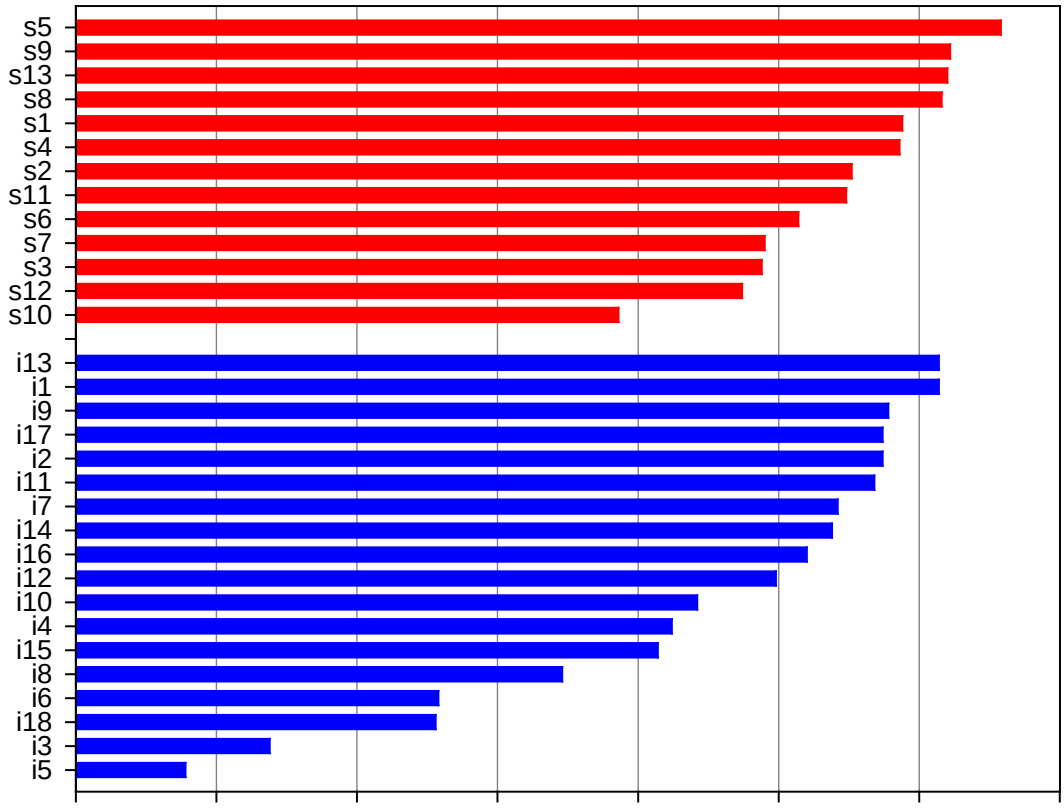


Figure L.2 represents the way in which adjustments could be made to the profile presented in Figure L.1 to more closely reproduce the precise pattern of scores across the 13 studies for a single criterion or across the 18 criteria for a single study. In other words, these graphic presentations represent a low-dimensional approximation of high-dimensional data by using three dimensions in the two figures combined to represent the variation between 18 criteria and 12 studies. (Reproducing the data exactly would likely take 13 dimensions.) The basis of the creation of the diagrams is that each successive dimension is chosen to maximize the amount of remaining variability that it can

account for. It is hoped that this form of data reduction can capture the principal variations among the items and the studies and that it might be possible to regard the "smaller" dimensions as being the result of random error; that is, inaccuracies in the reproduction of the actual data would be the result of random error that might not be present if the study were repeated. Whether the procedure provides a reasonable or useful reproduction in three dimensions is a judgment call.

Figure L.2: Second and third factors of variation (SigmaPlot® 11)

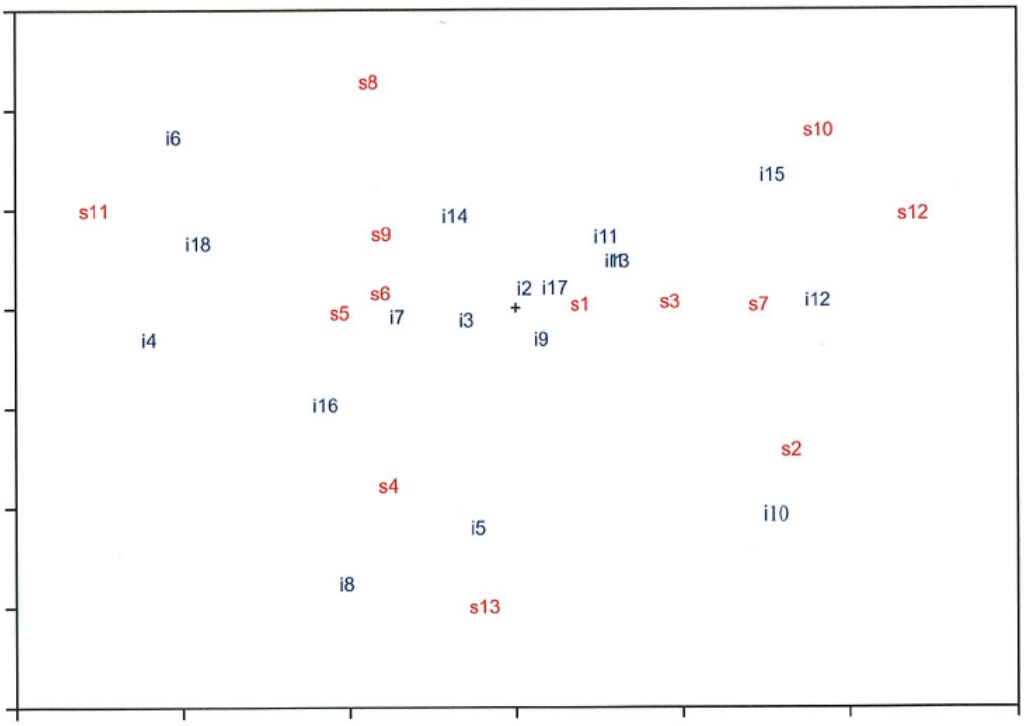
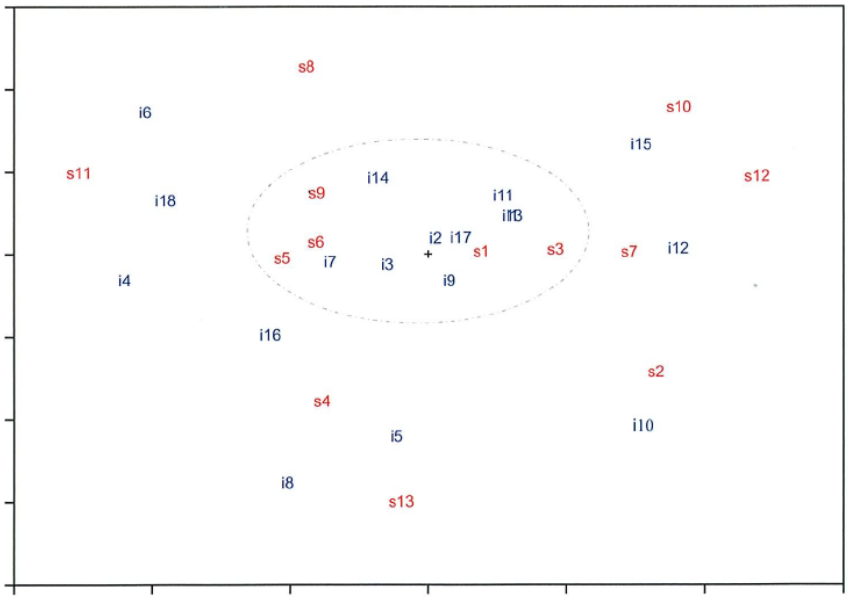


Figure L.2 is interpreted as follows.

1. Those criteria that are farthest from the origin (marked by "+") have a more deviant pattern, compared to the Figure L.1 across the studies than those criteria closest to the origin. Similarly, those studies that are farthest from the origin have a more deviant pattern, compared to the Figure L.1 across the items than those studies closest to the origin. We might conclude that criteria 2, 17, 9, 8, 3, 11, and 14 demonstrate a pattern across the studies as reflected in Figure L.1. Similarly, studies 1, 3, 5, 6, and 9 show a pattern across the items as reflected in Figure L.1 (Figure L.2.1).

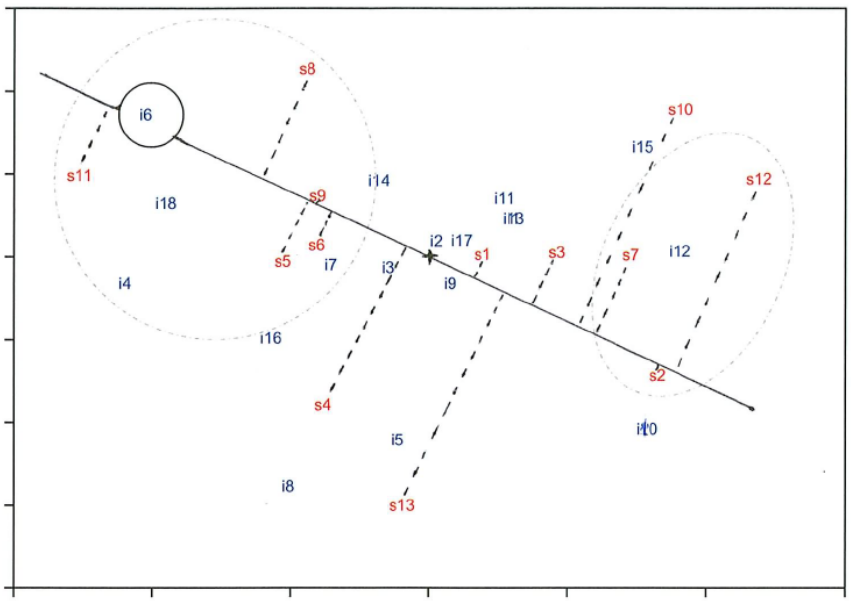
   For criteria/items and studies that have very high or very low mean scores, there is little variability "left over," which results in both in low standard deviations and in low 'scatter' among the profiles (i.e. closeness to the figure origin).

Figure L.2.1: Second and third dimensions of variation in the study by criterion/item matrix (SigmaPlot® 11)



2. For a deviant criterion (e.g. criterion i6) we can mentally draw a line from the i6 location through the "+" sign marking the origin. Now if we mentally draw lines from the studies perpendicular to this line, we will create an ordering of the studies (11, 8, 9, 5, 6, 4, 1, 13, 3, 10, 7, 12, 2). Those at the top (11, 8, 9, 5, 6) had higher scores on criterion 6 than Figure L.1 would have suggested; those at the bottom (7, 12, 2) had lower scores than Figure L.1 would have suggested (Figure L.2.2).

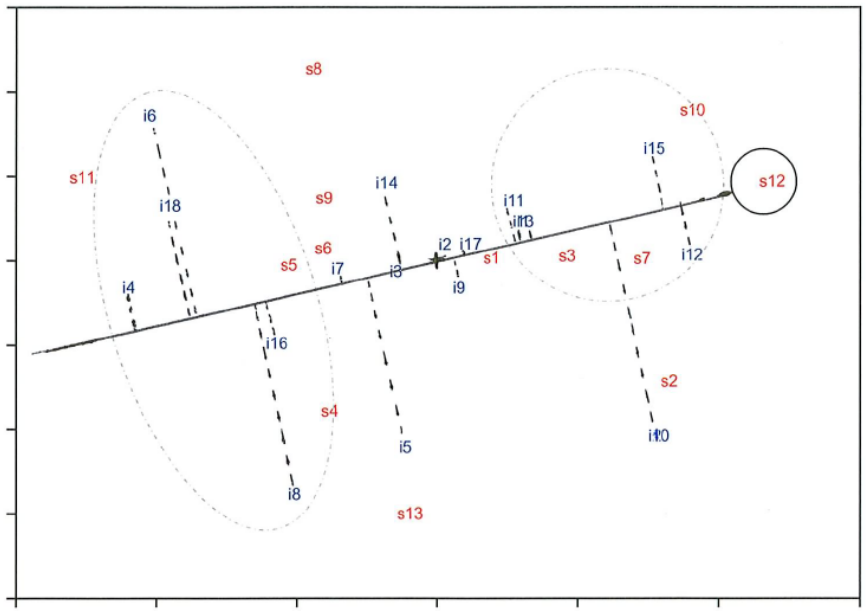
Figure L.2.2: Second and third dimensions of variation in the study by criterion/item matrix (SigmaPlot® 11)



3. For a deviant study (e.g. study 12), we can mentally draw a line from the s12 location through the "+" sign marking the origin. Now if we mentally draw lines from the studies perpendicular

to this line, we will create an ordering with criteria 15, 12, 11, 13 at the top of the ordering and criteria 4, 8, 16, 18, 6, 5 at the bottom. Analogous to the description above, this means that study 12 had relatively higher scores on criteria 15, 12, 11, and 13 than suggested by Figure L.1 whereas criteria 4, 8, 16, 18, 6, and 5 had lower scores (Figure L.2.3).

Figure L.2.3: Second and third dimensions of variation in the study by criterion/item matrix (SigmaPlot® 11)



4. Finally, studies that have small radial angles between them (e.g. 11, 8, 9, 6, 5) have similarly shaped patterns across the 18 criteria. Conversely, criteria with small radial angles between them (e.g. 16, 8, and 5) would tend to have similarly shaped patterns across the 13 studies. If we mentally drew a line from s10 through the origin and repeated the projections of the items, we would find that the projections provide an almost identical ordering to the one for study 12. This is because the two lines (the one through s12 and the origin, and the one through s10 to the origin) have a very small angle between them. If we were to repeat the exercise with s2, the angle between this line and the others would be greater, and the adjustments required of the profile in Figure L.1 would be different from those required for s10 and s12. That means that the prebuilt profiles for studies or items that are in the same quadrant of the figure will tend to look similar to each other (Figure L.2.4a and L.2.4b).

Figure L.2.4a: Second and third dimensions of variation in the study by criterion/item matrix (SigmaPlot® 11)

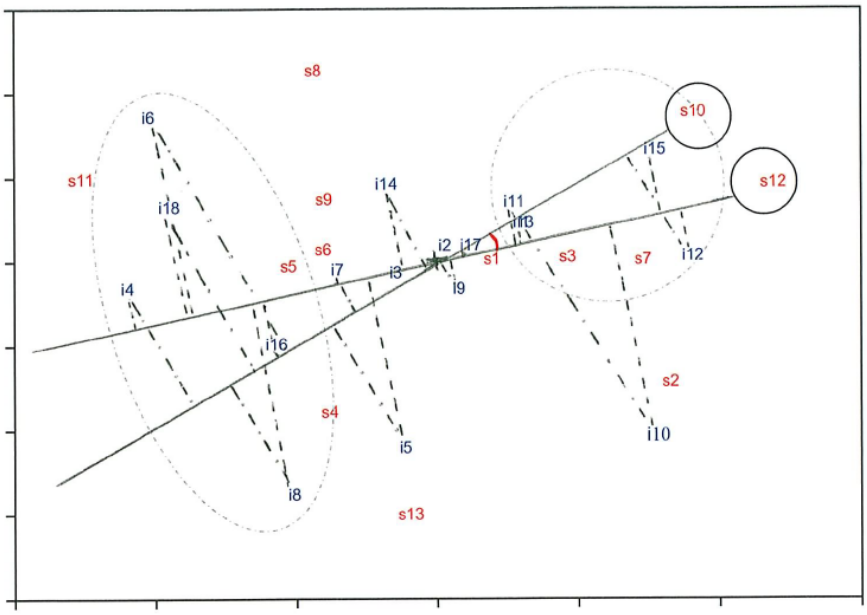
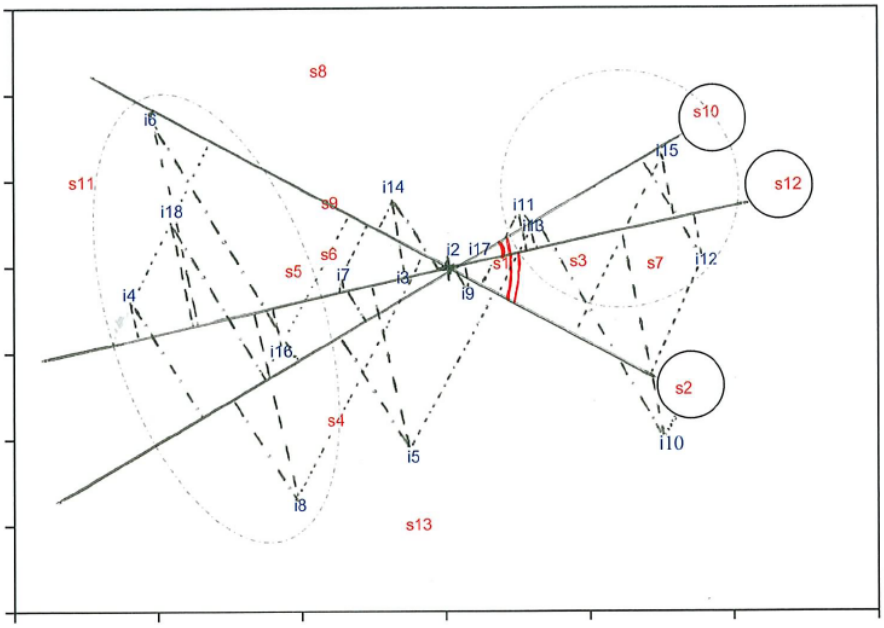


Figure L.2.4b: Second and third dimensions of variation in the study by criterion/item matrix (SigmaPlot® 11)



5. A close examination of these patterns can suggest which items might be structurally related across all studies, especially when the items are located far from the origin.

REFERENCES

1. Green S, Higgins J, editors. Glossary. In *Cochrane Handbook for Systematic Reviews of Interventions.* Edition 4.2.5. Cochrane Collaboration; 2005 May. Available: www.cochrane.org/resources/handbook/ (accessed 2009 Sep 11).
2. Dalziel K, Round A, Stein K, Garside R, Castelnuovo E, Payne L. *Do the findings of case series studies vary significantly according to methodological characteristics?* 2005. Available: www.hta.ac.uk/execsumm/summ902.htm (accessed 2009 Sep 11).
3. Stein K, Dalziel K, Garside R, Castelnuovo E, Round A. Association between methodological characteristics and outcome in health technology assessments which included case series. *International Journal of Technology Assessment in Health Care* 2005;21(3):277-87. Available: www.mrw.interscience.wiley.com/cochrane/ clcmr/articles/CMR-9844/frame.html (accessed 2009 Sep 11).
4. Mallen C, Peat G, Croft P. Quality assessment of observational studies is not commonplace in systematic reviews. *Journal of Clinical Epidemiology* 2006;59(8):765-9.
5. Saunders LD, Soomro GM, Buckingham J, Jamtvedt G, Raina P. Assessing the methodological quality of nonrandomized intervention studies. *Western Journal of Nursing Research* 2003;25(2):223-37.
6. Cauchi PA, Ang GS, Zuara-Blanco A, Burr JM. A systematic literature review of surgical interventions for limbal stem cell deficiency in humans. *American Journal of Ophthalmology* 2008;146(2):251-9.
7. Khan KS, ter Riet G, Glanville J, Sowden AJ, Kleijnen J, editors. *Understanding systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews* [01 April 2001]. Available: www.york.ac.uk/ inst/crd/pdf/crdreport4_complete.pdf (accessed 2009 Jul 28).
8. Huisstede B, Miedema HS, van Opstal T, de Ronde MT, Verhaar JA, Koes BW. Interventions for treating the radial tunnel syndrome: a systematic review of observational studies. *Journal of Hand Surgery American Volume* 2008;33A(1):72-8.
9. McCrory DC, Samsa GP, Hamilton BB, Govert JA, Matchar DB, Goslin RE, et al. Treatment of pulmonary disease following cervical spinal cord injury. *Report* 2001. Available: www.ahrq.gov/clinic/epcsums/spinalsum.htm.
10. Moga C, Harstall C. Gastric electrical stimulation (Enterra Therapy system) for the treatment of gastroparesis. *Report* 2006;73. Available: www.ihe.ca/documents/hta/HTA_Report_37.pdf.
11. National Collaborating Centre for Acute Care. Preoperative tests: the use of routine preoperative tests for elective surgery − appendices, guidelines and information: evidence, methods & guidance [2003 Jun 1:87-8]. Available: www.nice.org.uk/pdf/PreopTests_Apps.pdf (accessed 2009 Jul 31).
12. Taylor RS, Van Buyten JP, Buchser E. Spinal cord stimulation for complex regional pain syndrome: a systematic review of the clinical and cost-effectiveness literature and assessment of prognostic factors. *European Journal of Pain* 2006;10(2):91-101.

13. Yang AW. Assessing quality of case series studies: development and validation of an instrument by herbal medicine CAM researchers. *Journal of Alternative and Complementary Medicine* 2009;15(5):513-22.

14. Young J, Hyde C, Fry-Smith A, Gold L. Lung volume reduction surgery for chronic obstructive pulmonary disease with underlying severe emphysema. *Report* 1999;89. Available: http://rep.bham.ac.uk/pdfs/1999/lungvolreport.pdf.

15. Chipchase LS, Thoirs K, Jedrzejczak A. The effectiveness of real time ultrasound as a biofeedback tool for muscle retraining. *Physical Therapy Reviews* 2009;14(2):124-31.

16. American Academy for Cerebral Palsy and Developmental Medicine. *AACPDM methodology to develop systematic reviews of treatment interventions.* Revision 1.2. American Academy for Cerebral Palsy and Developmental Medicine; 2008. Available: www.aacpdm.org/publications/outcome/resources/ systematicReviewsMethodology.pdf (accessed 2009 Aug 13).

17. Bryant J, Cave C, Mihaylova B, Chase D, McIntyre L, Gerard K, et al. Clinical effectiveness and cost-effectiveness of growth hormone in children: a systematic review and economic evaluation. *Health Technology Assessment* 2002;6(18):1-168.

18. Nichol G, Stiell IG, Laupacis A, Pham B, De M, V, Wells GA. A cumulative meta-analysis of the effectiveness of defibrillator-capable emergency medical services for victims of out-of-hospital cardiac arrest. *Annals of Emergency Medicine* 1999;34(4 Part 1):517-25.

19. de Kleuver M, Oner FC, Jacobs WC. Total disc replacement for chronic low back pain: background and a systematic review of the literature. *European Spine Journal* 2003;12(2):108-16.

20. Des Jarlais DC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *American Journal of Public Health* 2004;94(3):361-6.

21. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology & Community Health* 1998;52(6):377-84.

22. Green C, Martin CW, Bassett K, Kazanjian A. A systematic review and critical appraisal of the scientific evidence on craniosacral therapy. *Report* 1999;54. Available: www.chspr.ubc.ca/node/373.

23. Haines SJ, Lapointe M. Fibrinolytic agents in the management of posthemorrhagic hydrocephalus in preterm infants: the evidence. *Childs Nervous System* 1999;15(5):226-34.

24. Hardy M, Coulter I, Venturupalli S, Roth EA, Favreau J, Morton SC, et al. Ayurvedic interventions for diabetes mellitus: a systematic review. *Report* 2001;155. Available: www.ahrq.gov/clinic/epcsums/ayurvsum.htm.

25. Hayashi M, Wilson NH, Yeung CA, Worthington H, V. Systematic review of ceramic inlays. *Clinical Oral Investigations* 2003;7(1):8-19.

26. Jongerius PH, van TP, van LJ, Gabreels FJ, Rotteveel JJ. A systematic review for evidence of efficacy of anticholinergic drugs to treat drooling. *Archives of Disease in Childhood* 2003;88(10):911-4. Available: http://adc.bmjjournals.com/cgi/content/ full/88/10/911.

27. MacDermid J. An introduction to evidence-based practice for hand therapists *Journal of Hand Therapy* 2004;17;105-17.

28. Merlin T, Arnold E, Petros P, Tulloch A, MacTaggart P, Jamieson G, et al. A systematic review of tension-free urethropexy for stress urinary incontinence: intravaginal slingplasty and the tension-free vaginal tape procedures. *Report* 2001;130. Available: onlinelibrary.wiley.com/doi/10.1046/j.1464-4096.2001.01667.x/full.

29. Mortenson PA, Eng JJ. The use of casts in the management of joint mobility and hypertonia following brain injury in adults: a systematic review. *Physical Therapy* 2003;83(7):648-58.

30. Oremus M, Zeidler J, Ensom MH, Matsuda-Abedini M, Balion C, Booker L, et al. *Utility of monitoring mycophenolic acid in solid organ transplant patients*. Rockville, MD: Agency for Healthcare Research and Quality; 2008 Feb. Evidence Report/Technology Assessment No. 164. AHRQ Publication No.08-E006. *Report* 2008;131.

31. Overend TJ, Anderson CM, Lucy SD, Bhatia C, Jonsson B, I, Timmermans C. The effect of incentive spirometry on postoperative pulmonary complications: a systematic review. *Chest* 2001;120(3):971-8. Available: www.chestjournal.org/ cgi/content/full/120/3/971.

32. PEDro scale. Available: http://fmweb01.ucc.usyd.edu.au/psycbite/ratings.shtml#2 (accessed 2009 Jul 29).

33. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (minors): development and validation of a new instrument.. *ANZ Journal of Surgery* 2003;73(9):712-6.

34. Stead LG, Gilmore RM, Bellolio MF, Rabinstein AA, Decker WW. Percutaneous clot removal devices in acute ischemic stroke: a systematic review and meta-analysis. *Archives of Neurology* 2008;65(8):1024-30.

35. Stewart A, Cummins C, Gold L, Jordan R, Phillips W. The effectiveness of the Mirena coil (levonorgestrel-releasing intrauterine system) in menorrhagia. *Report* 1999;34. Available: http://rep.bham.ac.uk/pdfs/1999/menorrhagia.pdf.

36. Thomas KC, Bailey CS, Dvorak MF, Kwon B, Fisher C. Comparison of operative and nonoperative treatment for thoracolumbar burst fractures in patients without neurological deficit: a systematic review. *Journal of Neurosurgery Spine* 2006;4(5):351-8.

37. Helm S, Hayek SM, Benyamin R, Manchikanti L. Systematic review of the effectiveness of thermal annular procedures in treating discogenic low back pain. *Pain Physician* 2009;12(1):207-32. Available: www.painphysicianjournal.com/2009/january/2009;12;207-232.pdf (accessed 2012 Feb 9).

38. Reiman MP, Harris JY, Cleland JA. Manual therapy interventions for patients with lumbar spinal stenosis: a systematic review. *New Zealand Journal of Physiotherapy* 2009;37(1):17-28. Available: www.highbeam.com/doc/1G1-207945598.html.

39. Smith TO, Hedges C, MacNair R, Schankat K. Early rehabilitation following less invasive surgical stabilisation plate fixation for distal femoral fractures. *Physiotherapy* 2009;95(2):61-75. Available: http://dx.doi.org/10.1016/j.physio.2009.02.002.

40. Deenadayalan Y, Perraton L, Machotka Z, Kumar S. Day therapy programs for adolescents with mental health problems: a systematic review. *Internet Journal of Allied Health Sciences and Practice* 2010;8(1):1-14. Available: http://ijahsp.nova.edu/articles/ Vol8Num1/kumar.htm.

41. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. 2009. Available: www.ohri.ca/ programs/clinical_epidemiology/oxford.htm (accessed 2009 Aug 19).

42. Cho MK, Bero L. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA* 1994;272(2):101-4.

43. Montoya A. Validation of the excited component of the positive and negative syndrome scale (PANSS-EC) in a naturalistic sample of 278 patients with acute psychosis and agitation in a psychiatric emergency room. *Health and Quality of Life Outcomes* 2011;9.

44. Spitzer WO Lawrence V, Dales R, Hill G, Archer MC, Clark P, et al. Links between passive smoking and disease; a best-evidence synthesis. *Clinical Investigative Medicine* 1990;(13);17-42.

45. Moher D, Jadad AR, Nichol G. Assessing the quality of randomized controlled trials: an annotated bibliography of checklists. *Controlled Clinical Trials* 1995;16:62-73.

46. Lionel NDW, Herxheimer A. Assessing reports of therapeutic trials. *BMJ* 1970;3:637-40.

47. Badgley RF. An assessment of research methods reported in 103 scientific articles from two Canadian medical journals. *CMAJ* 1961;85(5).

48. Thomson ME, Kramer MS. Methodological standards for controlled clinical trials of early contact and maternal-infant behaviour. *Pediatrics* 1984;73:294-300.

49. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analysis of randomised controlled trials. *N Engl J Med* 1987;316:450-5.

50. DerSimonian R, Charette J, McPeek B, Mosteller F. Reporting on methods in clinical trials. *N Engl J Med* 1982;306:1332-7.

51. Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials. *JAMA* 1994;272:1926-31.

52. Dalkey NC. The Delphi method: an experimental study of group opinion prepared for United States Air Force Project RAND. Santa Monica, CA: RAND Corporation, 1969. Available: 192.5.14.43/content/dam/rand/pubs/research_memoranda/ 2005/RM5888.pdf (accessed 2009 Jun 22).

53. Jones J, Hunter D. Consensus methods for medical and health services research. *BMJ* 1995; 311(7001):376-80.

54. Rowe G, Wright G, Bolger F. Delphi: A reevaluation of research and theory. *Technological Forecasting and Social Change* 1991;39(3):235-51.

55. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. *Systems to rate the strength of scientific evidence.* Rockville, MD: Agency for Healthcare Research and Quality; 2002Apr. Evidence Report/Technology Assessment No 47. AHRQ Publication No. 02-E016. Available: www.thecre.com/pdf/ahrq-system-strength.pdf 2002 (accessed 2011 Jul 7).

56. Elwood M, editor. *Critical appraisal of epidemiological studies and clinical trials.* 2nd edition. New York: Oxford University Press, 1998.

57. Matchar DB, Goldstein LB, McCrory DC, Oddone EZ, Jansen DA, Hilborne LH, Park RE. Carotid endarterectomy: a literature review and ratings of appropriateness and necessity. *RAND* 1992 (accessed 2009 Aug 24).

58. Guo B, Corabian P, Harstall C. *Islet transplantation for the treatment of type 1 diabetes: an update.* Edmonton, AB: Institute of Health Economics; 2008 Dec. Available: www.ihe.ca/publications/library/2008/islet-transplantation-for-the-treatment-of-type-1-diabetes---an-update/ (accessed 2010 Jun 25).

59. Lee TC, Barshes NR, O'Mahony CA, Nguyen L, Brunicardi FC, Ricordi C, et al. The effect of pancreatic islet transplantation on progression of diabetic retinopathy and neuropathy. *Transplant Proceedings* 2005;37(5):2263-5.

60. Froud T, Ricordi C, Baidal DA, Hafiz MM, Ponte G, Cure P, et al. Islet transplantation in type 1 diabetes mellitus using cultured islets and steroid-free immunosuppression: Miami experience. *American Journal of Transplantation* 2005;5(8):2037-46.

61. Ryan EA, Paty BW, Senior PA, Bigam D, Alfadhli E, Kneteman NM, et al. Five-year follow-up after clinical islet transplantation. *Diabetes* 2005;54(7):2060-9.

62. Hering BJ, Kandaswamy R, Ansite JD, Eckman PM, Nakano M, Sawada T, et al. Single-donor, marginal-dose islet transplantation in patients with type 1 diabetes. *JAMA* 2005;293(7):830-5.

63. Hering BJ, Kandaswamy R, Harmon JV, Ansite JD, Clemmings SM, Sakai T, et al. Transplantation of cultured islets from two-layer preserved pancreases in type 1 diabetes with anti-CD3 antibody. *American Journal of Transplantation* 2004;4(3):390-401.

64. Hirshberg B, Rother KI, Digon BJ, III, Lee J, Gaglia JL, Hines K, et al. Benefits and risks of solitary islet transplantation for type 1 diabetes using steroid-sparing immunosuppression: the National Institutes of Health experience. *Diabetes Care* 2003;26(12):3288-95.

65. Barshes NR, Vanatta JM, Mote A, Lee TC, Schock AP, Balkrishnan R, et al. Health-related quality of life after pancreatic islet transplantation: a longitudinal study. *Transplantation* 2005;79(12):1727-30.

66. Shapiro AMJ, Ricordi C, Hering BJ, Auchincloss H, Lindblad R, Robertson RP, et al. International trial of the Edmonton protocol for islet transplantation. *N Engl J Med* 2006;355(13):1318-30.

67. Maffi P, Bertuzzi F, De TF, Magistretti P, Nano R, Fiorina P, et al. Kidney function after islet transplant alone in type 1 diabetes: impact of immunosuppressive therapy on progression of diabetic nephropathy. *Diabetes Care* 2007;30(5):1150-5.

68. Venturini M, Fiorina P, Maffi P, Losio C, Vergani A, Secchi A, et al. Early increase of retinal arterial and venous blood flow velocities at color Doppler imaging in brittle type 1 diabetes after islet transplant alone. *Transplantation* 2006;81(9):1274-7.

69. O'Connell PJ, Hawthorne WJ, Holmes-Walker DJ, Nankivell BJ, Gunton JE, Patel AT, et al. Clinical islet transplantation in type 1 diabetes mellitus: results of Australia's first trial. *Medical Journal of Australia* 2006;184(5):221-5.

70. Keymeulen B, Gillard P, Mathieu C, Movahedi B, Maleux G, Delvaux G, et al. Correlation between beta cell mass and glycemic control in type 1 diabetic recipients of islet cell graft. *Proceedings of the National Academy of Sciences of the United States of America* 2006;103(46):17444-9.

71. Poggioli R, Ponte G, Baidal DA, Cure P, Pileggi A, Messinger S, et al. Quality of life after islet transplantation. *Diabetes* 2005;54 Suppl 1:A87.

72. Ebrahim S, Clarke M. STROBE: new standards for reporting observational epidemiology, a chance to improve. *International Journal of Epidemiology* 2007:1-3.

73. STROBE Initiative. STROBE statement: checklist of items that should be included in reports of observational studies. *International Journal of Public Health* 532008:3-4.

74. Fung AE, Palanki R, Bakri SJ, Depperschmidt E, Gibson A. Applying the CONSORT and STROBE statements to evaluate the reporting quality of neovascular age-related macular degeneration studies. *Ophthalmology* 2009;116(2):286-96.

75. Crombie I. *The pocket guide to critical appraisal: a handbook for heath care professionals*. London: BMJ, 1996.

76. Kumar S. *Scale to review case studies/series/reports*. Postgraduate Systematic Review thesis. Cape Town: University of South Australia, 1999.

77. Fleiss JL GA. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *Journal of Clinical Epidemiology* 1991;44:127-9.

78. Levine M, Walters S, Lee H, Haines T, Holbrook A, Moyer V, et al. User's guides to the medical literature IV: how to use an article about harm. *JAMA* 1994;271:1615-9.

79. How to read clinical journals IV: to determine etiology or causation. *CMAJ* 1981; 124:985-90.

80. van Tulder FA, Bombardier C, Bouter L. Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine* 2003:28(12):1290-9.

81. Lievense A, Bierma-Zeinstra S, Verhagen A, Verhaar J, Koes B. Influence of work on the development of osteoarthritis of the hip: a systematic review. *Journal of* R*heumatology* 2001;28:2520-8.

82. Borghouts JA, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. *Pain* 1998:77(1):1-13.

83. Hyde C, Wake B, Bryan S, Barton P, Fry-Smith A, Davenport C, et al. Fludarabine as second-line therapy for B cell chronic lymphocytic leukaemia: a technology assessment.

Health Technology Assessment 2002;6(2). Available: www.hta.ac.uk/fullmono/mon602.pdf (accessed 2009 Aug 14).

84. Wake B, Hyde C, Bryan S, Barton P, Song F, Fry-Smith A, et al. Rituximab as third-line treatment for refractory or recurrent Stage III or IV follicular non-Hodgkin's lymphoma: a systematic review and economic evaluation. *Health Technology Assessment* 2002;6(3):1-85. Available: www.hta.ac.uk/project.asp?PjtId=1293.

85. Law M, Stewart D, Pollock N, Letts L, Bosch J, Westmorland M. Critical review form: quantitative studies. McMaster University, 1998. Available: www.srs-mcmaster.ca/Portals/20/pdf/ebp/quanguidelines.pdf.

86. van Tulder MW, Assendelft WJJ, Koes BW, Bouter LM, the Editorial Board of the Cochrane Collaboration Back Pain Review Group. Method guidelines for systematic reviews in the Cochrane Collaboration back review group for spinal disorders. *Spine* 1997;22;2323-33.

87. Schechter M. Introduction to critical appraisal using the UBC Intervention Study Appraisal Form. Available: *members.shaw.ca/rbrands/ebm/therapeutic%20interventions/Schechter%20Introduction%20to%20Critical%20Appraisal.pdf* (accessed 2009 Jul 29).

88. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Deciding on the best therapy. In: *Clinical Epidemiology. A Basic Science for Clinical Medicine.* 2nd edition. Boston: Little, Brown, 1991:187-248.

89. Altman DG. Better reporting of randomized controlled trials: the CONSORT statement. *BMJ* 1996;(313);570-1.

90. Begg C, Moher D, Schulz K. Improving the quality of reporting the randomized controlled trials. *JAMA* 1996;(276);637-9.

91. Macfarlane TV, Glenny A-M, Worthington HV. Systematic review of population-based epidemiological studies of oro-facial pain. *Journal of Dentistry* 2001;(29);451-67.

92. Moher D, Shultz K, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Lancet* 2001;(357);1191-4.

93. Randall RC, Wilson NH. Glass-inomer restoratives: a systematic review of a secondary caries treatment effect. *Journal of Dentistry Research* 2009;(78);628-37.

94. Sackett DL, Department of Clinical Epidemiology and Biostatistics MUHC. How to read clinical journals, V: to distinguish useful from useless or even harmful therapy. *CMAJ* 1981;124:1156-62. Available: www.pubmedcentral.nih.gov/picrender.fcgi?artid=1705333&blobtype=pdf (accessed 2009 Aug 13).

95. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *International Journal of Epidemiology* 2007;36:666-76.

96. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. *International Journal of Qualitative Health Care* 162004:9-18.

97. van Tulder MW, Koes BW, Bouter LM. Conservative treatment of acute and chronic nonspecific low back pain: a systematic review of randomized controlled trials of the most common interventions. *Spine* 1997;22(18):2128-56.

98. Learning & Development Public Health Resource Unit. *CASP (Critical Skills Appraisal Programme)*. Oxford, UK: Learning & Development Public Health Resource Unit, 2007.

99. Flannery WH, Teukolsy BP, Vetterlin,WT. *Numerical recipes in C*. Cambridge, UK: Cambridge University Press, 1988.

Author Contribution Statements

Carmen Moga contributed to study conception and design, data analysis and interpretation, and approved the final version for publication.

Bing Guo contributed to study conception and design, data analysis and interpretation, and approved the final version for publication.

Don Schopflocher contributed to data analysis and interpretation and approved the final version for publication.

Christa Harstall contributed to study conception and design, revision of manuscript for critical content, and approved the final version for publication.

This Methodology Paper summarizes the process, a modified Delphi approach, used to develop a specific checklist for the quality appraisal of case series studies. This work was supplemented with a review of other published checklists and an initial pilot test of the newly developed quality appraisal checklist. Researchers at the Institute of Health Economics with researchers from two other Health Technology Assessment agencies in Australia and Spain were actively engaged throughout the Delphi process.

IHE INSTITUTE OF HEALTH ECONOMICS ALBERTA CANADA

Institute of Health Economics
1200 – 10405 Jasper Avenue
Edmonton AB Canada T5J 3N4
Tel. 780.448.4881 Fax. 780.448.0018
info@ihe.ca

www.ihe.ca

ISBN 978-1-926929-04-0 (on-line)