General Section
Research Paper

OPEN

PAIN REPORTS®

Content validation of a Critical Appraisal Tool for Reviewing Analgesia Studies (CATRAS) involving subjects incapable of self-reporting pain

Leon N. Warne[a,b,*], Stephan A. Schug[c,d], Thierry Beths[b], Juliana T. Brondani[e], Jennifer E. Carter[b], B. Duncan X. Lascelles[f,g,h,i], Anthea L. Raisis[a], Sheilah A. Robertson[j], Paulo V.M. Steagall[k,l], Polly M. Taylor[m], Ted Whittem[b], Sébastien H. Bauquier[b]

Abstract

Introduction: This article reports the content validation of a Critical Appraisal Tool designed to Review the quality of Analgesia Studies (CATRAS) involving subjects incapable of self-reporting pain and provide guidance as to the strengths and weakness of findings. The CATRAS quality items encompass 3 domains: level of evidence, methodological soundness, and grading of the pain assessment tool.

Objectives: To validate a critical appraisal tool for reviewing analgesia studies involving subjects incapable of self-reporting pain.

Methods: Content validation was achieved using Delphi methodology through panel consensus. A panel of 6 experts reviewed the CATRAS in 3 rounds and quantitatively rated the relevance of the instrument and each of its quality items to their respective domains.

Results: Content validation was achieved for each item of the CATRAS and the tool as a whole. Item-level content validity index and kappa coefficient were at least greater than 0.83 and 0.81, respectively, for all items except for one item in domain 2 that was later removed. Scale-level content validity index was 97% (excellent content validity).

Conclusions: This 67-item critical appraisal tool may enable critical and quantitative assessment of the quality of individual analgesia trials involving subjects incapable of self-reporting pain for use in systematic reviews and meta-analysis studies.

Keywords: Critical appraisal tool, Analgesia, Pain

1. Introduction

The scientific literature contains many examples of inconsistencies regarding the analgesic efficacy of treatments for pain. It is the authors' opinion that many of these inconsistencies arise due to variations in the quality and rigour with which each study was designed.

Determining the quality of published studies of analgesic interventions is difficult, particularly those involving animal pain models, hampering efforts to draw meaningful clinical conclusions from published findings and to perform systematic reviews of the literature.[34,40] Frameworks are needed to ensure greater consistency in experimental design to allow for more accurate comparison of findings.[40] The Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) has developed consensus reviews and recommendations for

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

[a] School of Veterinary and Life Sciences, Murdoch University, Perth, Western Australia, Australia, [b] Melbourne Veterinary School, The University of Melbourne, Werribee, Victoria, Australia, [c] Discipline of Anaesthesiology and Pain Medicine, Medical School, University of Western Australia, Perth, Western Australia, Australia, [d] Department of Anaesthesia and Pain Medicine, Royal Perth Hospital, Perth, Western Australia, Australia, [e] Universidade Estadual Paulista, Botucatu, Brazil, [f] Department of Clinical Sciences, Comparative Pain Research and Education Centre, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, USA, [g] Comparative Medicine Institute, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, USA, [h] Center for Pain Research and Innovation, UNC School of Dentistry, Chapel Hill, NC, USA, [i] Department of Anesthesiology, Center for Translational Pain Research, Duke University, Durham, NC, USA, [j] Lap of Love Veterinary Hospice and In-Home Euthanasia, Lutz, FL, USA, [k] Department of Clinical Sciences, Faculty of Veterinary Medicine, Université de Montréal, Saint-Hyacinthe, QC, Canada, [l] Animal Pharmacology Research Group of Quebec (GREPAQ), Faculty of Veterinary Medicine, Université de Montréal, Saint-Hyacinthe, QC, Canada, [m] Taylor Monroe, Ely, United Kingdom

**Corresponding author. Address: Veterinary Anaesthesia and Pain Management, College of Veterinary Medicine, School of Veterinary and Life Sciences, Murdoch University, 90 South St, Murdoch, Western Australia 6150, Australia. Tel.: +61 8 9360 6823. E-mail address: Leon.Warne@murdoch.edu.au (L.N. Warne).*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.painrpts.com).

PR9 3 (2018) e670

http://dx.doi.org/10.1097/PR9.0000000000000670

improving the design, execution, and interpretation of clinical analgesia trials in self-reporting humans.[9,33,48] Despite this, recommendations for animals, as well as for humans incapable of self-reporting pain are still lacking.

Descriptors of quality currently used to critically appraise scientific literature typically include schemes for assessing the *level of evidence* (LOE) and *methodological soundness*.[2,14] The LOE for a particular study is assigned according to the study design and its inherent likelihood to exclude bias. Grading of the methodological soundness of a study is typically based on how closely it conforms to established standards for study design. The "2011 Levels of Evidence" established by the Oxford Centre for Evidence-Based Medicine (OCEBM) and the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system are 2 of the most universally recognised ranking systems.[7,17,19,43] These critical appraisal tools (CATs) include checklists with specific questions and/or scales for scoring components of quality, which are combined to give a summary score.[24]

Current "gold-standard" pain assessment tools (PATs) rely on self-reporting, requiring an individual subject to both process external information and communicate this personal experience. In relation to subjects incapable of self-reporting pain such as noncommunicative or cognitively impaired human patients or pain in animal, this is not possible. In these situations, the assessment of pain involves changes in behavioural or physiological parameters. However, their use can be associated with considerable shortcomings. They may be unreliable, hampered by observational bias, or influenced by disease processes or pharmacological interventions. Current evidence indicates that treatment of pain is inadequate in human patients incapable of self-reporting pain as well as animal, largely because of inadequate methods of pain assessment.[23,27,32,39,44,52] This brings into question the accuracy and validity of findings from published studies using non–self-reporting PATs. Several assessment techniques have been designed specifically for the *grading of PATs* in noncommunicative human patients; these primarily evaluate the psychometric properties of PATs against: item selection and content validation, reliability, validity, feasibility, and relevance or impact on patient outcomes.[16,38,52]

Critical appraisal and interpretation of findings from analgesia studies remains challenging because of the lack of a single CAT for evaluating all domains (ie, *LOE*, *methodological soundness*, and *grading of the PATs*). The aim of the present work was to construct and validate a CAT that incorporated the aforementioned 3 domains to assess the quality of individual analgesia trials and provide quantification of quality for use in systematic reviews and meta-analysis studies focusing on subjects incapable of self-reporting pain. Importantly, this CAT will assess whether the methodologies used by a study conform to appropriately high scientific standards, independent of the species being studied. This study reports the content validation of this CAT referred to as "CATRAS" for Critical Appraisal Tool for Reviewing Analgesia Studies.

2. Methods

A working group identified and adapted potential domains and items to form the preliminary version of the CATRAS. The weight of each domain and associated item scores were assigned by the working group at this stage. This study reports the initial validation of the CATRAS by a panel of experts and comprised 2 phases: (1) development of consensus and preliminary agreement of content and (2) content validation. An overview of these phases is represented in **Figure 1**.

2.1. Identification and adaption of the domains and items

In 2014, the working group that comprised 2 of the authors (L.N. W. and S.H.B.) identified 3 domains, which would form the analytical scope of the CATRAS, and ultimately the framework used to critically appraise published analgesia studies.

(1) *Level of evidence* (CATRAS step 1, domain 1)—The working group adopted without modification, an LOE classification system that was previously used by a landmark systematic review initiative, the Reassessment Campaign on Veterinary Resuscitation (RECOVER). The LOE classification used by the RECOVER initiative was itself modified from a major human review group, the 2010 International Liaison Committee on Resuscitation (ILCOR 2010).[4,31] This domain contained 6 items (LOE 1–6), which are characterised by criteria presented in **Table 1**, domain 1. The LOE of a study must be established before assessment of its methodological soundness.

(2) *Methodological soundness* (CATRAS step 2, domain 2)—The list of quality items contributing to this domain was adopted with minor modifications from that used in the RECOVER initiative process, which was originally derived from CATs designed by the OCEBM.[4,6] Modifications included the addition of the following 2 quality item questions to each of the 5 possible categories (A–E): "*Was conflict of interest stated?*" and, "*Was the statistical methodology of the study appropriate? (If "NO," please justify).*"

(3) *Grading of the PAT* (CATRAS step 3, domain 3)—The purpose of domain 3 is to provide critical appraisal of the PAT used in a study being evaluated. To assess the quality of the PAT used within an analgesia study being reviewed, domain 3 of the CATRAS requires the investigator to review the original or revised literature describing the development, refinement, or validation of the PAT. To achieve the third domain of the CATRAS, a CAT was developed based on a psychometric scoring system developed and validated to evaluate PATs used in noncommunicative critically ill human patients.[16] The original psychometric scoring system used by Gélinas (2013) incorporated the GRADE system methodology.[16,21] The content of the psychometric scoring system used by Gélinas (2013) was adopted largely unmodified from the original version, with only the following minor changes being made by the working group: substitution of the word "*scale*" for "*Pain Assessment Tool*" OR "*PAT*" to maintain continuity with nomenclature in other domains; all references to "*ICU patients*" were removed to maintain relevance to more than just ICU patients; anthropological examples used within the scoring legend of items within the original tool were substituted for species' neutral descriptors; a minor change was also made to the terminology of the question in item 3.5 from "*Discriminant validation*" to "*Sensitivity to change*" as the working group considered that this change provided greater clarity.

2.2. Recruitment of the panel of experts

A panel of 6 internationally recognised experts from different institutions was established to critique the CATRAS. The expert panel comprised 6 of the authors (T.B., J.T.B., B.D.X.L., S.A.R., P.V.M.S., and P.M.T.). Panel members were selected based on their professional certifications and credentials, clinical experience and publication profile in translational research, veterinary pain management, and PAT construction.

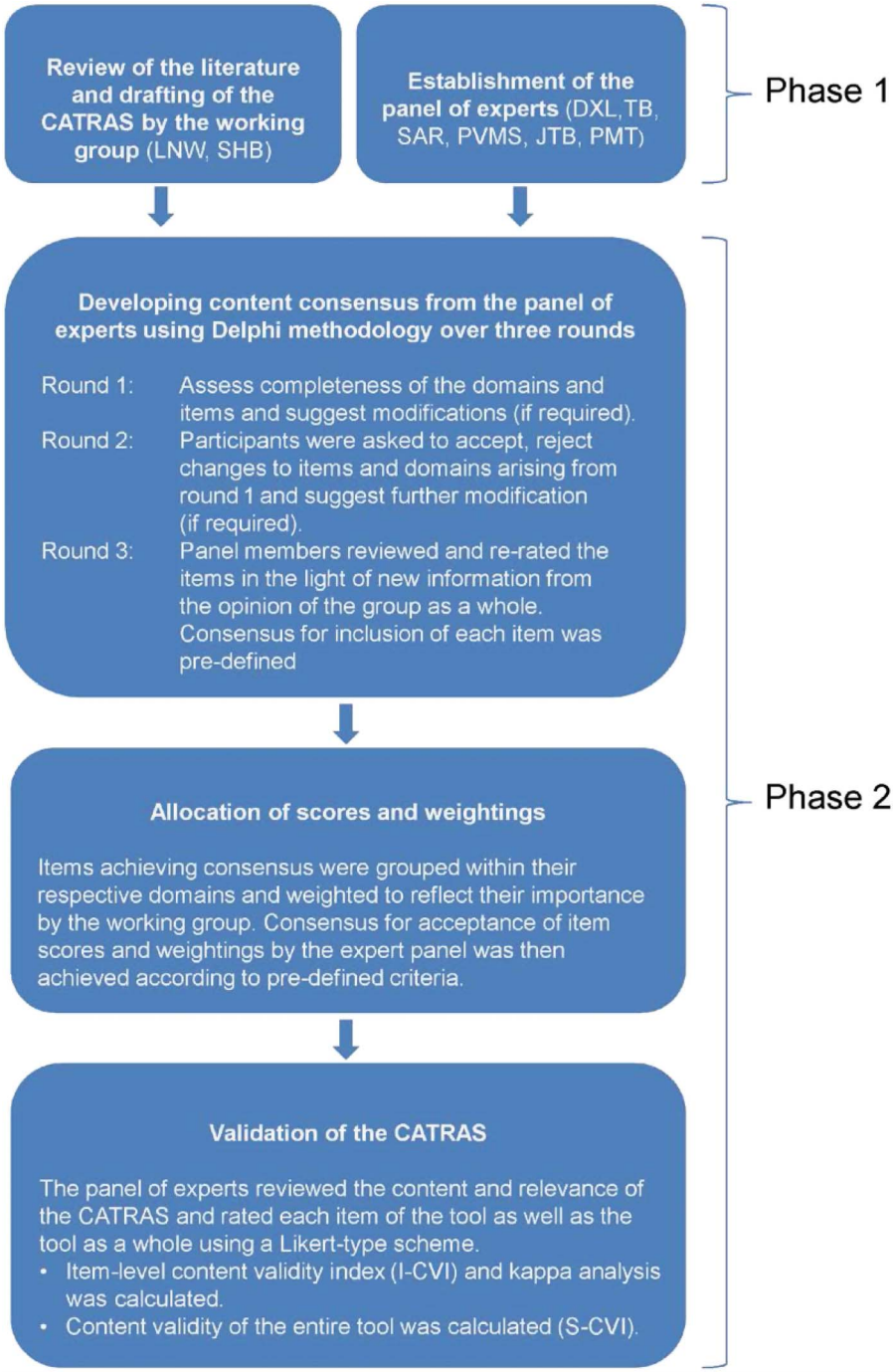

Figure 1. Diagrammatical representation of the sequence of tasks in this study. S-CVI, scale-level content validity index. LNW, Leon N. Warne; SHB, Sébastien H. Bauquier; DXL, B. Duncan X. Lascelles; TB, Thierry Beths; SAR, Sheilah A. Robertson; PVMS, Paulo V.M. Steagall; JTB, Juliana T. Brondani; PMT, Polly M. Taylor.

2.3. Development of consensus

The study used Delphi methodology to develop consensus from a panel of experts, by means of surveys conducted over 3 rounds, to ensure that the 3 domains and items generated in the development of the CATRAS were not merely a function of the smaller working group by means of surveys conducted over 3 rounds.[22,29] The objective of the first round was to gauge the completeness of the domains (and items within each) to assess adequately the quality of analgesia studies. Definitions for the domains and items within each were provided to enable comparison of each domain and item against its definition. Members of the expert panel were invited to contribute as many ideas as they wished in response to 2 open-ended questions regarding quality in analgesia studies: (1) "Are there additional domains beyond those already encompassed (ie, *LOE*, *methodological soundness*, and *grading of the PAT*), which you consider integral in comprehensively assessing the quality of analgesia studies? If YES, please list and explain your answer." (2) Within the existing 3 domains, what factors not already encompassed by existing items (if any), do you consider important for assessing the quality of analgesia studies?"

In the second round, the responses obtained in round 1 were collated into one document by the working group and redistributed

Table 1

CATRAS quality items and associated item-level content validity index (I-CVI) values and kappa coefficients (κ).

Domain 1: level of evidence (LOE)				
Study characteristics	**LOE**	**I-CVI**	**κ**	**Evaluation**
Randomised negative- or positive-controlled trials (RCTs), or meta-analyses of RCTs in the target species: Clinical studies that prospectively collect data and randomly allocate the subjects to intervention or control groups, or meta-analyses of these studies.	1	1	1	Excellent
Prospective clinical studies in the target species using concurrent controls (ie, controls recruited at the same time as experimental subjects) without randomisation. These studies can be: 1. Interventional clinical: Include subjects who are allocated to intervention or control groups concurrently, but in a nonrandom fashion OR 2. Observational clinical: Include cohort and case–control studies.	2	1	1	Excellent
Experimental laboratory study in the target species: These could include, but are not limited to randomized, blinded, and controlled studies.	3	1	1	Excellent
Clinical retrospective studies in the target species: The study and control groups have been selected from a previous period in time (historical controls).	4	0.83	0.81	Excellent
Case series and case reports in the target species: A single group of subjects exposed to the intervention (or factor under study), but without a control group.	5	1	1	Excellent
Studies, experimental or clinical, that are not directly related to the specific target species or target population. These could be different species/populations, including experimental models in nontarget species.	6	0.83	0.81	Excellent
Domain 2: methodological soundness		**I-CVI**	**κ**	**Evaluation**
[A] Randomised controlled trials				
Was the randomisation list concealed?		1	1	Excellent
Were all subjects who entered the trial accounted for at its conclusion?		1	1	Excellent
Were the subjects analysed in the groups to which they were randomised?		0.83	0.81	Excellent
Were owners, investigators, and evaluators "blinded" to which treatment was being received?		1	1	Excellent
Aside from the experimental treatment, were the groups treated equally?		0.83	0.81	Excellent
Were the groups similar at the start of the trial?		1	1	Excellent
Is the relevance to the question being posed high (ie, no confounding factors such as concomitant drug administration or intervention, which could bias the relevance to the question, the study directly addresses the question)?		1	1	Excellent
Is there a high likelihood that the administered drug (or intervention) will have clinically relevant analgesic effect?		0.83	0.81	Excellent
Was conflict of interest stated?		1	1	Excellent
Was ethical/institutional review board approval of the study stated?		1	1	Excellent
Was the statistical methodology of the study appropriate? If "NO," please justify:		1	1	Excellent

(*continued on next page*)

Table 1 (*continued*)

CATRAS quality items and associated item-level content validity index (I-CVI) values and kappa coefficients (κ).

Domain 2: methodological soundness	I-CVI	κ	Evaluation
[B] Clinical studies using concurrent controls without randomisation			
Were comparison groups clearly defined?	1	1	Excellent
Were outcomes measured in the same (preferably blinded) objective way in both groups?	1	1	Excellent
Were known confounders identified and appropriately controlled for?	1	1	Excellent
Was follow-up of subjects sufficiently long and complete?	1	1	Excellent
Is the relevance to the question being posed high?	1	1	Excellent
Is there a high likelihood that the administered drug (or intervention) will have clinically relevant analgesic effect?	1	1	Excellent
Was conflict of interest stated?	1	1	Excellent
Was ethical/institutional review board approval of the study stated?	1	1	Excellent
Was the statistical methodology (including sample size) of the study appropriate? If "NO," please justify:	1	1	Excellent
[C] Experimental laboratory studies in target species			
Is the study a randomised controlled trial (RCT)?	1	1	Excellent
Is the study a randomised trial (including RCT)?	1	1	Excellent
Is the relevance to the question being posed high?	1	1	Excellent
Is there a high likelihood that the administered drug (or intervention) will have clinically relevant analgesic effect?	1	1	Excellent
Was conflict of interest stated?	1	1	Excellent
Was ethical/institutional review board approval of the study stated?	1	1	Excellent
Was the statistical methodology (including sample size) of the study appropriate? If "NO," please justify:	1	1	Excellent
[D] Retrospective clinical studies using controls without randomisation			
Were comparison groups clearly defined?	0.83	0.81	Excellent
Were outcomes measured in the same objective way in both groups?	0.83	0.81	Excellent
Were known confounders identified and appropriately controlled for?	1	1	Excellent
Was follow-up of subjects sufficiently long available and complete?	1	1	Excellent
Were criteria used to include subjects in or exclude subjects from the study clearly stated?	1	1	Excellent
Is the relevance to the question being posed high?	1	1	Excellent
Is there a high likelihood that the administered drug (or intervention) will have clinically relevant analgesic effect?	1	1	Excellent
Was conflict of interest stated?	1	1	Excellent
Was ethical/institutional review board approval of the study stated?	0.83	0.81	Excellent
Was the statistical methodology (including sample size) of the study appropriate? If "NO," please justify:	1	1	Excellent
[E] Clinical studies without controls			
Were outcomes measured in an objective way?	1	1	Excellent
Were known confounders identified and appropriately controlled for?	1	1	Excellent
Was follow-up of subjects sufficiently long and complete?	1	1	Excellent
Is the relevance to the question being posed high?	1	1	Excellent

(*continued on next page*)

Table 1 (*continued*)

CATRAS quality items and associated item-level content validity index (I-CVI) values and kappa coefficients (κ).

Domain 2: methodological soundness	I-CVI	κ	Evaluation
Is there a high likelihood that the administered drug (or intervention) will have clinically relevant analgesic effect?	1	1	Excellent
Was conflict of interest stated?	1	1	Excellent
Was ethical/institutional review board approval of the study stated?	1	1	Excellent
Was the statistical methodology (including sample size) of the study appropriate? If "NO," please justify:	1	1	Excellent

Domain 3: grading of the pain assessment tool (PAT)					
Question		**Scoring legend**	**I-CVI**	**κ**	**Evaluation**
1.1	Does the PAT assess multiple important indicators or dimension of pain? Nb: Each tool receives 2 points if it contains both *psychomotor/visual assessment* of pain (eg, posture, comfort, activity, and demeanour) and *interactive assessment* of pain (eg, response to palpation of potential pain loci), 1 point, if it contained only *psychomotor/visual assessment* or only *interactive assessment* of pain, and 0 points, if it did not cover either of these dimensions/ indicators of content validity.	2: PAT covers all important items or dimensions 1: PAT covers important items or dimensions to a moderate extent 0: PAT does not seem to cover important items or dimensions	1	1	Excellent
1.2	Was the process of item selection described?	2: PAT was developed for a specific population, using a theoretical or conceptual framework, or a qualitative approach was used (eg, consultation with clinicians) 1: PAT was developed based on literature review only 0: No information is provided about item selection	1	1	Excellent
1.3	Was content evaluated by experts (content validation)?	2: Content was evaluated by experts in the field, and CVI was calculated for each item included in the PAT 1: Content was evaluated by experts, but no CVI is reported 0: No information is provided about content validation	0.83	0.81	Excellent
1.4	Are limitations of some items presented or discussed?	1: No limitations or if any limitations, they are presented and item modifications have been made or precautions have been stated 0: No information is provided	1	1	Excellent
2.1	Was internal consistency (Cronbach's α coefficient) of the PAT calculated?	2: $0.70 < \alpha < 0.90$ 1: $0.60 < \alpha \leq 0.70$ or $\alpha \geq 0.90$ 0: $\alpha \leq 0.60$ or no information provided	1	1	Excellent
2.2	Was interrater reliability (Cohen's kappa coefficient if quantitative) calculated?	2: kappa >0.60 or intraclass correlation coefficient (ICC) > 0.80 1: 0.60 ≥ kappa >0.40 or 0.60 < ICC ≤0.80 0: kappa ≤0.40, ICC ≤0.60 or no information provided	1	1	Excellent
2.3	Was interrater reliability tested with other raters besides the research team?	1: Other raters than the research staff members were involved 0: Only research staff Members were involved	0.83	0.81	Excellent
2.4	Was intrarater reliability tested? Optional—to be examined if ICC <0.80 for interrater reliability.	2: kappa >0.60 or ICC >0.80 1: 0.60 ≥ kappa >0.40 or 0.60 < ICC ≤ 0.80 0: kappa ≤0.40, ICC ≤0.60 or no information provided	1	1	Excellent
3.1	What is the total of participants for the purpose of testing the PAT?	2: N > 50 1: 20 < N ≤ 50 0: N ≤ 20	0.83	0.81	Excellent

(*continued on next page*)

Table 1 (*continued*)

CATRAS quality items and associated item-level content validity index (I-CVI) values and kappa coefficients (κ).

Domain 3: grading of the pain assessment tool (PAT)					
Question		**Scoring legend**	**I-CVI**	**κ**	**Evaluation**
3.2	Criterion validation: Was the PAT correlated with the current "gold standard" or with a measure renowned in the field of interest if no "gold standard" has been established?	2: $r > 0.60$ with the comparison measure 1: $0.40 < r \leq 0.60$ 0: $r \leq 0.40$ or no information provided	1	1	Excellent
3.3	Criterion validation: Was the sensitivity of the PAT calculated?	2: Sensitivity ≥ 80% 1: 60% ≤ sensitivity < 80% 0: Sensitivity <60% or no information provided	1	1	Excellent
3.4	Criterion validation: Was the specificity of the PAT calculated?	2: Specificity ≥ 80% 1: 60% ≤ specificity < 80% 0: Specificity <60% or no information provided	1	1	Excellent
3.5	Sensitivity to change: Was the PAT able to differentiate between different situations (eg, between pain and no pain; before and after the administration of an analgesic; changes in health status of the patient)?	2: A significant difference was found 1: A difference was found but was not significant 0: No difference was found or no information is provided	1	1	Excellent
4.1	Is the PAT easily applied to the clinical setting?	1: PAT is short and manageable 0: PAT is more complex or no information is provided	1	1	Excellent
4.2	Are directives of use of the PAT clearly described?	1: Yes, directives of use including the scoring method are described 0: No information about directives of use is provided	1	1	Excellent
5.1	Was the relevance of the PAT or impact of its implementation in patient outcomes examined?	1: PAT is considered to be useful and relevant to practice by more than 80% of clinicians; use of the PAT yielded a significant change into practice (eg, better use of medication and increase in patients' assessments) 0: PAT is not considered to be useful/relevant to practice by more than 20% of clinicians; use of the PAT did not yield a significant change in practice or no information provided	1	1	Excellent

Evaluation criteria for kappa used guidelines described in Cicchetti and Sparrow (1981) and Fleiss et al. (2013): *Fair* = kappa of 0.40 to 0.59; *Good* = kappa of 0.60 to 0.74; and *Excellent* = kappa > 0.74.[8,15]

to the panel for individual rating of relevance as well as evaluation of the clarity of item construction and wording. Participants were asked to *accept*, *reject*, or *suggest modification to* points arising from round 1 relating to existing items or suggest additional items within each domain.

Information acquired from round 2 was then incorporated into the round 3 questionnaires with the addition of the participant's own ratings and comments for each item as a reminder. Thus, separate round 3 questionnaires were developed for each member of the expert panel. Panel members reviewed and rerated the items in the light of new information from the opinion of the group as a whole. Consensus for inclusion of each item was predefined as acceptance by 4 or more members of the expert panel (>4/6, >66%) without any further modification being recommended by any of the endorsing members. Any modifications were to be rated individually in a subsequent round, with consensus for inclusion being predefined as previously described.

2.4. Content validation

The panel of experts reviewed the content and relevance of the CATRAS and evaluated the appropriateness of each item of the tool as well as the tool's relevance as a whole using the following Likert-type scheme: 1 = not relevant, 2 = somewhat relevant, 3 = quite relevant, and 4 = very relevant.[28]

2.4.1. Item-level content validity index and kappa analysis

For each item of the CATRAS, the content validity (item-level content validity index [I-CVI]) was calculated by dividing the number of experts assigning a rating of either 3 or 4 by the total number of experts—that is, the proportion of experts in agreement concerning the relevance. For example, an item rated as "quite relevant" or "very relevant" by 4 of 6 experts would have an I-CVI of 0.67.[37] The kappa coefficient for individual items was also calculated using previously described methodology.[50] Evaluation criteria for kappa used guidelines described in Cicchetti and Sparrow (1981) and Fleiss et al. (2013): *Fair* = kappa of 0.40 to 0.59; *Good* = kappa of 0.60 to 0.74; and *Excellent* = kappa >0.74.[8,15] Items were considered to have adequate content validity for inclusion in the CATRAS if they achieved an I-CVI of 0.83 or greater and a kappa coefficient of 0.81 or greater. Kappa coefficients and I-CVI were calculated, and based on published recommendations, a cutoff point for an item to remain in the tool was predefined as 0.81 and 0.83, respectively (reflecting one disagreement).[30,37]

2.4.2. Scale-level content validity index

The content validity of the tool as a whole (scale-level CVI [S-CVI]) was evaluated using previously described methodology, whereby the S-CVI is calculated as the average I-CVI across all items of the tool.[37] Based on published recommendations, the minimum S-CVI required for the CATRAS to achieve content validity for the tool as a whole was predefined as 0.90.[37,49]

2.5. Development of the quantitative aspects of the assessment tool (CATRAS)

A final list of quality items that achieved consensus agreement from the panel of experts was collated by the working group. Items were grouped within their respective domains and weighted to reflect their importance within each domain.

(1) Domain 1—LOE: The highest score of 6 was attributed to the strongest LOE (LOE 1) and the lowest score of 1 to the weakest LOE (LOE 6) (Supplemental Table, available at http://links.lww.com/PR9/A29).
(2) Domain 2—Methodological soundness: The LOE of a study (step 1) must be established before assessment of its methodological soundness (Supplemental Table, available at http://links.lww.com/PR9/A29). Domain 2 consists of a list of methodological quality items (step 2 categories A–E) relating to the LOE assigned in step 1 (Supplemental Table, available at http://links.lww.com/PR9/A29). The overall methodological soundness is defined by 3 quality terms, as either: "good," "fair," or "poor." Studies are to be assigned the quality-term "good" if they contain most or all the relevant quality items, "fair" if they contain some of the relevant quality items, and "poor" if they contain only a few of the relevant quality items but were considered to be of sufficient value to warrant inclusion in the next step of the review. The 3 quality terms "good," "fair," and "poor" were given a score of 3, 2, and 1, respectively.
(3) Domain 3—Grading of the PAT: Items were assigned weighted scores representing those ascribed by Gélinas.[16]

The sum total of all weighted items within each individual domain was transcribed into a percentage by dividing the attributed score by the maximum possible score of each respective domain. Transcription of the score into a percentage allowed for standardisation between the 3 domains of the CATRAS. Each domain was assigned equal weighting. Members of the expert panel were then asked to *accept*, *reject*, or *suggest modification* to the allocation of item scores and weightings. Consensus for inclusion of item scores and weightings were predefined as acceptance by 4 or more members of the expert panel without any further modification being recommended by any of the endorsing members. Any modifications were to be rated individually in a subsequent round, with consensus for inclusion being predefined as previously described.

3. Results

3.1. Evaluation for completeness and development of consensus

During the round 1 review of the draft CATRAS, no additional domains were deemed necessary by any member of the expert panel. The following item was added to the methodological soundness domain in each of the 5 possible categories (A–E): "*Was ethical/institutional review board approval of the study stated?*" This addition was unanimously accepted during rounds 2 and 3 of the review process, and no further items were modified or excluded.

3.2. Content validation

3.2.1. Item-level content validity index and kappa analysis

The 67 items of the 3 domains were reviewed (**Table 1**). Fifty-seven (57/67; 85%) items received 100% agreement by all 6 members of the expert panel (I-CVI = 1; kappa coefficient = 1). Ten (10/67; 13%) items received 83% agreement (I-CVI = 0.83; kappa coefficient = 0.81).

3.2.2. Scale-level content validity index

The content validity of the final remaining 67 items of the CATRAS resulted in a 97% (S-CVI = 0.97) agreement, indicating that the tool achieved *excellent* content validity.

3.3. Development of the quantitative aspects of the assessment tool (CATRAS)

After content validation, the working group assigned scores and weightings to the 67 quality items and the associated domains. The assigned scores and weightings for all 67 quality items and the associated domains were unanimously accepted without modification by the expert panel. The final results derived from application of the CATRAS are 3% scores (ie, one for each domain). The final version of the CATRAS is shown in the Supplemental Table (available at http://links.lww.com/PR9/A29). An example of the application of the CATRAS can be found in (Supplemental Appendix 1, http://links.lww.com/PR9/A22).

4. Discussion

In 2014, a working group (L.N.W. and S.H.B.) found no evidence of a published CAT designed specifically to evaluate the quality of published analgesia studies in any species. To address this absence, the authors designed and validated a 67-item CAT (Supplemental Table, available at http://links.lww.com/PR9/A29) that incorporates 3 domains (*LOE*, *methodological soundness*, and *grading of the PAT*).

Level of evidence is used by many review processes to create order and simplicity from the heterogeneity of published studies, and is assigned according to the study type and its inherent likelihood to exclude bias.[4,31] The methodological quality and transparent reporting of an analgesia study is a key factor to consider when assessing its translational value.[36,41] The quality items listed within the CATRAS to assess methodological soundness are primarily based on those used in the RECOVER initiative process, which were originally derived from CATs designed by the OCEBM.[4,6] In addition, all the quality items described in methodological soundness category A of the CATRAS are part of the Consolidated Standards of Reporting Trials (CONSORT) 2010 "checklist of information to include when reporting a randomised trial."[35] The CONSORT statement was developed to improve the standard of reporting of randomised controlled trials for medical interventions.[3] Furthermore, the methodological soundness domain of the CATRAS also complies with the Animal Research Reporting *In Vivo* Experiments (ARRIVE) guidelines methodology section, which highlights details of bias reduction tactics such as sample size calculation, random allocation to groups, and observer blinding.[25,41] After recommendations from the panel of experts, the item "was

ethical/institutional review board approval of the study stated?" was added to the methodological soundness domain (categories A, B, C, D, and E) both to strengthen the tool and to promote ethical research.

Content validity concerns the degree to which a scale has an appropriate sample of items to represent the construct of interest; that is, whether the domain(s) of content for the construct is adequately represented by the items.[37] Content validity of the CATRAS was reviewed using a panel of 6 experts selected according to previously defined criteria.[12,18] In addition, widespread geographical distribution of the panel members (Australia, Brazil, Canada, United Kingdom, and United States) allowed for differences in colloquial terms that could affect instrument comprehension by many diverse groups.[18]

There are potential biases in the methodology used. First, the statements are not an inventory of every aspect of methodology that could impact on trial quality. In an attempt to reduce the likelihood of this bias, the working group obtained consensus opinion from individuals with direct experience of conducting studies involving assessment of pain; as knowledge of the subject matter is considered the most significant assurance of a valid outcome using the Delphi methodology.[47] Second, the reliability of the findings relating to validity coefficients may have been influenced by including the same individuals in both the content consensus and subsequently also as raters during the content validity process. The working group attempted to minimise this bias by both sequencing the order of the tasks and by their temporal separation: the consensus process occurred approximately 10 months before the content validity process.

A widely accepted method of quantifying content validity for multi-item tools such as CATRAS is the CVI based on expert rating of relevance.[37] The CVI is an index of consensus and the extent to which experts share a common interpretation of the construct of a tool.[46] A CVI was calculated for each quality item of the 3 domains (I-CVI) as well as for the overall CATRAS as the whole tool (S-CVI), thereby providing an index of interrater agreement. Critics of the CVI cite concerns about the possibility of inflated values because of the risk of chance agreement.[50] In an attempt to address this issue, the current study used a previously described modified kappa-like index that adjusts each I-CVI for chance agreement or disagreement.[37] Fifty-seven items received 100% agreement (I-CVI = 1); and ten items received 83% agreement (I-CVI = 0.83); there was no consistency observed in relation to individual members of the expert panel who rejected items of the CATRAS (including geographical location of the experts). Based on previously published guidelines citing the acceptable I-CVI in relation to the number of expert raters, these items of the CATRAS achieved adequate item-level content validity. In addition, using previously described evaluation criteria for kappa, these items were considered to have *excellent* agreement on relevance.[8,15]

Assessment of the S-CVI for the combined final 67 items of the CATRAS resulted in a 97% (0.97) agreement, indicating that the tool achieved *excellent* content validity. This result is considerably higher than published recommendations of the minimum S-CVI required for validation. Critical appraisal tool developers often set a criterion of 80% (0.80) or better agreement among expert reviewers as the lower limit of acceptability for an S-CVI.[11] However, we chose to adopt the more stringent recommendations of Waltz et al.[49] who set the lower limit of acceptable agreement at 90% (0.90).

It is widely accepted that interpretation of the results of a particular study should be informed by the quality of all aspects of the trial: the higher the quality, the greater the confidence in the validity and utility of the findings.[51] When evaluating analgesia studies, the quality of the PAT used must be considered as a significant factor determining the quality of the report. Previously, consideration of the impact of the PAT on the strength of evidence has not been possible. Development of the CATRAS may now enable evaluation of the strength of findings from published analgesia studies based on a more thorough assessment of quality. The need to assess the original or revised literature for the purpose of grading the PAT may decrease the time efficiency of the CATRAS and might be considered a limitation by some users. Future work could streamline this process by establishing precalculated grades for the commonly used PATs. Another potential limitation of the CATRAS could be that because of the nonlinearity of both visual analogue scales and numeric rating scales, calculation of sensitivity and specificity (domain 3, item 3.3 and 3.4) may not be relevant for assessing these scales and as such they may be slightly downgraded by up to 4% of the overall possible score for domain 3. An interpretation of the results obtained from the grading of a PAT can be found in (Supplemental Appendix 2, http://links.lww.com/PR9/A23).

In evaluating the quality of published analgesia studies, it is important to consider both the analgesic efficacy and the safety of the drugs and regimens used.[9] The CATRAS does not assess patient safety outcomes, and this may be considered a limitation.

In recent years, there has been growing interest in assessment of pain in animal subjects as well as in human patients incapable of self-reporting, with the number of newly developed PATs growing rapidly.[5,10,13,26,27,38,45] It is the responsibility of researchers, funding agencies, and journals to prevent excessive growth of nonvalidated PATs in circumstances where appropriate and validated tools already exist. Further psychometric evaluation of existing PATs should be given priority over developing new tools for future use. Valid, practical, and reliable PATs can add to the body of knowledge about pain and help improve its treatment.

The final results derived from application of the CATRAS are 3% scores (ie, one for each domain). Illustration of these results should be at the discretion of the investigators; however, a radar chart would allow for a two-dimensional representation of the 3% scores on axes starting from the same point (Supplemental Appendix 1, http://links.lww.com/PR9/A22). Charting of the data in this way enables clear visual representation of the quality of a specific analgesia study in relation to the *LOE*, *methodological soundness*, and *grading of the PAT*.

Central to the practice of evidence-based medicine (EBM) is the process of asking well-focused questions, searching for the best available evidence, critically appraising that evidence for quality and validity, then applying the results to improve patient outcomes.[20,42] The CATRAS is designed to facilitate the practice of EBM by enabling a quantitative quality assessment of an individual published study's evidence supporting (or rejecting) the clinical question being investigated. The CATRAS can be used to explore the influence of study quality or design methodology on the strength of the results and conclusions. We endorse the use of a systematic EBM approach that provides an explicit framework for formulating the clinical question or statement under investigation in terms of its 4 key parts—Problem/Population, Intervention, Comparison, and Outcome (PICO).[1]

The CAT developed in this study offers several benefits for assessing the quality of analgesia studies involving subjects incapable of self-reporting pain: its content was developed through the consensus of experts; it captures features of study design methodology, which are widespread in this field; and

content validation has been established. The next step in the development of this important tool would be to apply the CATRAS in a systematic review of the literature focused on questions arising from analgesia studies.

Disclosures

The authors have no conflict of interest to declare.

Acknowledgements

The authors thank Associate Professor Graham Hepworth of The University of Melbourne, Statistical Consulting Centre, for his statistical advice.

Appendix A. Supplemental digital content

Supplemental digital content associated with this article can be found online at http://links.lww.com/PR9/A29, http://links.lww.com/PR9/A22, and http://links.lww.com/PR9/A23.

Article history:
Received 3 May 2018
Received in revised form 8 June 2018
Accepted 11 June 2018

References

[1] Armstrong EC. The well-built clinical question: the key to finding the best evidence efficiently. WMJ 1999;98:25–8.
[2] Atkins D, Eccles M, Flottorp S, Guyatt GH, Henry D, Hill S, Liberati A, O'Connell D, Oxman AD, Phillips B, Schünemann H, Edejer TTT, Vist GE, Williams JW; GRADE Working Group. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches the GRADE Working Group. BMC Health Serv Res 2004;4:38.
[3] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. JAMA 1996;276:637–9.
[4] Boller M, Fletcher DJ. RECOVER evidence and knowledge gap analysis on veterinary CPR. Part 1: evidence analysis and consensus process: collaborative path toward small animal CPR guidelines. J Vet Emerg Crit Care 2012;22:S4–S12.
[5] Brondani JT, Mama KR, Luna SPL, Wright BD, Niyom S, Ambrosio J, Vogel PR, Padovani CR. Validation of the English version of the UNESP-Botucatu multidimensional composite pain scale for assessing postoperative pain in cats. BMC Vet Res 2013;9:143.
[6] CEBM. Oxford Centre for evidence-based medicine–critical appraisal worksheets (English). Available at: http://www.cebm.net/critical-appraisal/2005. Accessed June 2, 2015.
[7] CEBM. Oxford Centre for evidence-based medicine–levels of evidence (March 2009). Available at: http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/2009. Accessed June 2, 2015.
[8] Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic 1981;86:127–37.
[9] Cooper S, Desjardins P, Turk D, Dworkin R, Katz N, Kehlet H, Ballantyne J, Burke L, Carragee E, Cowan P, Croll S, Dionne R, Farrar J, Gilron I, Gordon D, Iyengar S, Jay G, Kalso E, Kerns R, McDermott M, Raja S, Rappaport B, Rauschkolb C, Royal M, Segerdahl M, Stauffer J, Todd K, Vanhove G, Wallace M, West C, White R, Wu C. Research design considerations for single-dose analgesic clinical trials in acute pain: IMMPACT recommendations. PAIN 2016;157:288–301.
[10] Corbett A, Achterberg W, Husebo B, Lobbezoo F, de Vet H, Kunz M, Strand L, Constantinou M, Tudose C, Kappesser J, de Waal M, Lautenbacher S. An international road map to improve pain assessment in people with impaired cognition: the development of the Pain Assessment in Impaired Cognition (PAIC) meta-tool. BMC Neurol 2014;14:229.
[11] Davis L. Instrument review: getting the most from a panel of experts. Appl Nurs Res 1992;5:194–97.
[12] Davis L, Grant J. Guidelines for using psychometric consultants in nursing studies. Res Nurs Health 1993;16:151–55.
[13] de Grauw JC, van Loon JPAM. Systematic pain assessment in horses. Vet J 2016;209:14–22.
[14] Dixon-Woods M, Booth AJ, Sutton AJ. Synthesizing qualitative research: a review of published reports. Qual Res 2007;7:375–422.
[15] Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. New Jersey: John Wiley & Sons, 2013.
[16] Gélinas C. A validated approach to evaluating psychometric properties of pain assessment tools for use in nonverbal critically ill adults. Semin Respir Crit Care Med 2013;34:153.
[17] Gordon HG, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schünemann HJ. Rating quality of evidence and strength of recommendations: GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–26.
[18] Grant J, Davis L. Selection and use of content experts for instrument development. Res Nurs Health 1997;20:269–74.
[19] Group GW. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490–94.
[20] Guyatt G, Cairns J, Churchill D, Cook D, Haynes B, Hirsh J, Irvine J, Levine M, Levine M, Nishikawa J, Sackett D, Brill-Edwards P, Gerstein H, Gibson J, Jaeschke R, Kerigan A, Neville A, Panju A, Detsky A, Enkin M, Frid P, Gerrity M, Laupacis A, Lawrence V, Menard J, Moyer V, Mulrow C, Links P, Oxman A, Sinclair J, Tugwell P. Evidence-based medicine: a new approach to teaching the practice of medicine. JAMA 1992;268:2420–25.
[21] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck Ytter Y, Alonso Coello P, Schünemann HJ. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–26.
[22] Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000;32:1008–15.
[23] Hellyer PW. Treatment of pain in dogs and cats. J Am Vet Med Assoc 2002;221:212–15.
[24] Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 2001;323:42–6.
[25] Kilkenny C, Browne W, Cuthill I, Emerson M, Altman D. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol 2010;8:e1000412.
[26] Langford D, Bailey A, Chanda M, Clarke S, Drummond T, Echols S, Glick S, Ingrao J, Klassen Ross T, LaCroix Fralish M, Matsumiya L, Sorge R, Sotocinal S, Tabaka J, Wong D, van den Maagdenberg AMJM, Ferrari M, Craig K, Mogil J. Coding of facial expressions of pain in the laboratory mouse. Nat Methods 2010;7:447–49.
[27] Lichtner V, Dowding D, Esterhuizen P, Closs SJ, Long A, Corbett A, Briggs M. Pain assessment for people with dementia: a systematic review of systematic reviews of pain assessment tools. BMC Geriatr 2014;14:138.
[28] Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;140:1–55.
[29] Linstone HA, Turoff M. The Delphi method: techniques and applications. Reading: Addison-Wesley Publishing Company, 1975.
[30] Lynn MR. Determination and quantification of content validity. Nurs Res 1986;35:382–5.
[31] Martín-Hernández H, López-Messa JB, Pérez-Vela JL, Herrero-Ansola P. ILCOR 2010 recommendations. The evidence evaluation process in resuscitation. Med Intensiva 2011;35:249–55.
[32] Marzinski LR. The tragedy of dementia: clinically assessing pain in the confused, nonverbal elderly. J Gerontol Nurs 1991;17:25–8.
[33] McGrath P, Walco G, Turk D, Dworkin R, Brown M, Davidson K, Eccleston C, Finley GA, Goldschneider K, Haverkos L, Hertz S, Ljungman G, Palermo T, Rappaport B, Rhodes T, Schechter N, Scott J, Sethna N, Svensson O, Stinson J, von Baeyer C, Walker L, Weisman S, White R, Zajicek A, Zeltzer L. Core outcome domains and measures for pediatric acute and chronic/recurrent pain clinical trials: PedIMMPACT recommendations. J Pain 2008;9:771–83.
[34] Mignini LE, Khan KS. Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research. BMC Med Res Methodol 2006;6:10.
[35] Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux P, Elbourne D, Egger M, Altman DG. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869–c69.
[36] Percie du Sert N, Rice ASC. Improving the translation of analgesic drugs to the clinic: animal models of neuropathic pain. Br J Pharmacol 2014;171:2951–63.
[37] Polit D, Beck C, Owen S. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health 2007;30:459–67.

[38] Pudas-Tähkä SM, Axelin A, Aantaa R, Lund V, Salanterä S. Pain assessment tools for unconscious or sedated intensive care patients: a systematic review. J Adv Nurs 2009;65:946–56.
[39] Resnik D, Rehm M. The undertreatment of pain: scientific, clinical, cultural, and philosophical factors. Med Health Care Philos 2001;4: 277–88.
[40] Rice AS, Cimino-Brown D, Eisenach JC, Kontinen VK, Lacroix-Fralish ML, Machin I, Mogil JS, Stöhr T, Consortium PP. Animal models and the prediction of efficacy in clinical trials of analgesic drugs: a critical appraisal and call for uniform reporting standards. PAIN 2008;139:243–47.
[41] Rice ASC, Morland R, Huang W, Currie G, Sena E, Macleod M. Transparency in the reporting of in vivo pre-clinical pain research: the relevance and implications of the ARRIVE (animal research: reporting in vivo experiments) guidelines. Scand J Pain 2013;4:58–62.
[42] Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak 2007;7:16.
[43] Schünemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, Fahy BF, Gould MK, Horan KL, Krishnan JA, Manthous CA, Maurer JR, McNicholas WT, Oxman AD, Rubenfeld G, Turino GM, Guyatt G. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. Am J Respir Crit Care Med 2006;174:605–14.
[44] Sengstaken EA, King SA. The problems of pain and its detection among geriatric nursing home residents. J Am Geriatr Soc 1993;41:541–44.
[45] Sotocinal SG, Sotocina S, Sorge R, Zaloum A, Tuttle A, Martin L, Wieskopf J, Mapplebeck J, Wei P, Zhan S, Zhang S, McDougall J, King O, Mogil J. The Rat Grimace Scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. Mol Pain 2011;7:55.
[46] Stemler SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Pract Assess Res Eval 2004;9:1.
[47] Stone F, Busby D. The Delphi research methods in family therapy. New York: Guildford, 1996.
[48] Turk D, Dworkin R, Burke L, Gershon R, Rothman M, Scott J, Allen R, Atkinson JH, Chandler J, Cleeland C, Cowan P, Dimitrova R, Dionne R, Farrar J, Haythornthwaite J, Hertz S, Jadad A, Jensen M, Kellstein D, Kerns R, Manning D, Martin S, Max M, McDermott M, McGrath P, Moulin D, Nurmikko T, Quessy S, Raja S, Rappaport B, Rauschkolb C, Robinson J, Royal M, Simon L, Stauffer J, Stucki G, Tollett J, von Stein T, Wallace M, Wernicke J, White R, Williams A, Witter J, Wyrwich K. Developing patient-reported outcome measures for pain clinical trials: IMMPACT recommendations. PAIN 2006;125:208–15.
[49] Waltz CF, Strickland OL, Lenz ER. Measurement in nursing and health research. New York: Springer Publishing Company, 2010.
[50] Wynd CA, Schmidt B, Schaefer MA. Two quantitative approaches for estimating content validity. West J Nurs Res 2003;25:508–18.
[51] Yates S, Morley S, Eccleston C, de C Williams AC. A scale for rating the quality of psychological trials for pain. PAIN 2005;117:314–25.
[52] Zwakhalen SM, Hamers JP, Abu-Saad HH, Berger MP. Pain in elderly people with severe dementia: a systematic review of behavioural pain assessment tools. BMC Geriatr 2006;6:3.