Send Orders of Reprints at reprints@benthamscience.net

The Open Orthopaedics Journal, 2013, *7*, (Suppl 4: M7) 461-472

Open Access

A Description of the Methodology Used in an Overview of Reviews to Evaluate Evidence on the Treatment, Harms, Diagnosis/Classification, Prognosis and Outcomes Used in the Management of Neck Pain

P. Lina Santaguida[*,1], Homa Keshavarz [2], Lisa C. Carlesso[3], Margaret Lomotan[3], Anita Gross[2], Joy C. MacDermid[4], David M. Walton[5] and ICON Working Group[§]

[1]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada*

[2]*Department of Onclology, McMaster University, Hamilton, Ontario, Canada*

[3]*School of Rehabilitation Sciences, McMaster University, Hamilton, Ontario, Canada*

[4]*School of Rehabilitation Sciences, McMaster University, Hamilton Ontario Canada and Co-Director, Clinical Research Lab, Hand and Upper Limb Centre, St. Joseph's Health Centre, London, Ontario, Canada*

[5]*School of Physical Therapy, Western University, London, Ontario, Canada*

Abstract: *Background*: Neck Pain (NP) is a common musculoskeletal disorder and the literature provides conflicting evidence about its management.

Objective: To describe the methodology used to conduct an overview of reviews (OvR) and to characterize the distribution and risk of bias profiles across the evidence for all areas of NP management.

Methods: Standard systematic review (SR) methodology was employed. MEDLINE, CINAHL, EMBASE, ILC, Cochrane CENTRAL, and LILACS were searched from 2000 to March 2012; Narrative and SR and clinical practice guidelines (CPG) evaluating the efficacy of treatment (benefits and harms), diagnosis/classification, prognosis, and outcomes were eligible. For treatment, articles were limited to SRs from 2005 forward. Risk of bias of SR was assessed with the AMSTAR; the AGREE II was used to critically appraise the CPGs.

Results: From 2476 articles, 508 were eligible for full text screening. A total of 341 articles were included. Treatment (n=117) had the greatest yield. Other clinical areas had less literature (diagnosis=54, prognosis=16, outcomes=27, harms=16). There were no SR for classification and narrative reviews were problematic for this topic. There was great overlap across different databases within each clinical area except for those for outcome measures. Risk of bias assessment using the AMSTAR of eligible SRs showed a similar trend across different clinical areas.

Conclusion: A summary of methods used to review the literature in five clinical areas of NP management have been described. The challenges of selecting and synthesizing eligible articles in an OvR required customized solutions across different areas of clinical focus.

Keywords: Evidence syntheses, overview of reviews, neck pain.

INTRODUCTION

The evidentiary base in some areas of musculoskeletal problems is sufficiently mature that a substantive number of systematic reviews (SRs) exist, particularly when the topic includes numerous interventions and patient groups. An overview of reviews (OvR) (often labeled as umbrella reviews or review of reviews) is one approach to summarizing large amounts of evidence by collating findings from a series of SR [1]. OvR can be described as a relatively new approach to summarizing large amounts of evidence from SR in a single useful report, especially in areas where there are overlapping SR. Ideally, a SR will evaluate all relevant literature within a specific population for either interventions or other areas related to the management of the problem. Typically, SR will be limited to either a narrowly defined population or to a single intervention or comparator. Evaluation of a single SR may not provide a more global perspective that considers all populations likely to benefit from the intervention or compare across different interventions [2]. Additionally, SR evaluating the same or similar topics may present conflicting results; understanding why conclusions differ (or not) across SR and the implications of the discordance is becoming increasingly important for end users (clinicians, guideline developers, policymakers). One approach to address these issues is with

*Address correspondence to this author at the Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada; Tel: 905-525-9140; Fax: 905-522-7681;
E-mail: santag@mcmaster.ca

[§]ICON is a multi-disciplinary collaborative group that includes scientist-authors (listed below) and support staff (Margaret Lomotan) that conduct knowledge synthesis and translation aimed at reducing the burden of neck pain.

the undertaking of an OvR. Using methods similar to traditional SR, OvR attempt to identify high-quality, relevant SR and evaluate the consistency of findings across different reviews, thereby taking a more comprehensive approach to evaluating findings across different interventions, outcomes, adverse events, and patient populations [1, 2]. In some clinical areas, there are numerous existing SR and replication of already existing work within similar SR may not be efficient or ethical from a resource perspective [3]; undertaking an OvR would detect such overlap as well as areas where there are gaps. All these reasons reflect the growing need and increasing trend towards undertaking OvR that serve to provide a panorama view of evidence, particularly when the clinical questions are multifaceted and complex or summary across clinical management areas are considered.

In the context of managing neck pain (NP), an OvR would be ideal to provide an assessment of the evidentiary base across different clinical areas. NP has a high prevalence and has been associated with considerable disability and cost; for these reasons NP is considered an important musculoskeletal problem, particularly in industrialized countries [4]. NP can cause differing levels of pain and disability, resulting in impaired daily function, headaches, nerve-related problems (often radiating from the neck to the shoulder, arm, and hand) or cognitive disturbances [4]. The widespread frequency and recurrence of NP has led to hundreds of primary studies and reviews (both narrative and systematic) evaluating all aspects of clinical management of NP, including treatment, harms (associated with treatment), diagnosis, prevention, prognosis, and outcomes. The large and ever increasing literature base has provided a rich field of inquiry across which findings can be justifiably summarized.

This paper summarizes the methods used to prepare several OvR across five key clinical areas (treatment, harms, diagnosis/classification, prognosis, and outcomes) to assess the distribution and quality of the evidentiary basis. By providing evidence summaries in these key areas we hope to promote discussion and facilitate eventual agreement in the research community, and to inform future evidence-based recommendations for International Guidelines on the management of NP. The detailed findings for each area of clinical management of NP (treatment, harms, diagnosis, prognosis, and outcomes) are presented elsewhere (some in this same issue). However, in addition to describing the OvR methods used, the purpose is also to describe the relative distribution of the NP literature and the risk of bias profiles **across** the different areas of clinical management, which is one advantage of undertaking an OvR of this scope.

METHODS

Standard SR methodology was employed with respect to searching for relevant articles and for SR selection, extraction, and risk of bias assessment. A literature search was undertaken from January 2000 to March 2012. Eligible articles included SR and clinical practice guidelines (CPG) for all five clinical areas of NP management. We included narrative reviews, consensus statements or commentaries for NP classification, diagnosis, prognosis, and outcomes. Not being certain of the number of relevant SR available, we hypothesized that key information might be available in narrative reviews (in the absence of SR) with respect to NP classification systems and outcome measures; we anticipated that these clinical areas may have less SR and also may be poorly indexed (for example, classification systems). For articles evaluating the efficacy of treatment in NP, we restricted the eligible studies to SR published from 2005 forward due to the large volume of eligible reviews. This approach to restricting by year of publication for areas of greater activity is consistent with methods proposed when there are multiple similar SR; in this way the most chronologically recent and methodologically rigourous reviews would be screened when there is overlap with older SR [5].

Types of Participants in the Reviews and CPG

Articles had to include populations that had any type of NP to be eligible. The anatomical boundaries of the neck are defined as commencing from the occiput (base of the skull) to the upper thoracic spine (T1-T6, mid upper back) and can include the upper regions of the torso or shoulder area. The upper shoulder region is included within the definition because muscles attached to the neck are also attached as far as the lateral ends of the scapulae. Some articles included populations with general musculoskeletal or chronic pain that could potentially include NP; from these studies, only those that reported stratified data for NP, or some primary studies with NP or that included some primary studies with greater than 50% of NP patients in the sample, were eligible. Studies were limited to those that included adult participants (18 years or older) who complained of NP.

All types of non-malignant NP were eligible including both non-specific (no specific identifiable etiology) and specific groups. Non-specific NP is considered to have a postural or mechanical basis and includes whiplash associated disorders (WAD I to II), myofascial neck pain, degenerative changes, and cervicogenic headache (corresponding to diagnostic classification 11.2.1 proposed by the International Headache Society classification and diagnostic criterion for headache disorders) [6] Different types of non-specific NP are subdivided on the basis of etiology into non-traumatic and traumatic (e.g. whiplash). Fracture, dislocation, neurological disorders and fibromyalgia associated with the neck were excluded from non-specific neck pain. Specific NP includes radicular pain (including WAD III), cervical disc prolapse pain, and facet joint pain.

Exclusions: Articles were excluded if they investigated NP with the following: a) trauma associated with fracture or head injury; b) definite or possible long tract neurological signs (e.g., myelopathies); c) NP caused by other pathological entities (e.g., tumours, infections); d) headache not of cervical origin but associated with the neck (e.g., migraine, tension-type headache).

Types of Interventions in the Reviews and CPG

For SR evaluating the efficacy of treatment or harms associated with the treatment, the interventions were limited to manual therapy, physical medicine modalities, drug therapies, patient education and communication, ergonomics,

orthoses, prevention, and psychological interventions. For SR evaluating treatment and harms, only reviews that had searched at least two bibliographic databases and had undertaken some form of assessment of risk of bias of the included studies were eligible; this restriction was not applied to the other clinical areas (prognosis, diagnosis, or outcomes). This approach is consistent with methods for OvR to ensure only the best evidence is included in the findings and to minimize duplication [5]. Exclusions: treatment SR published prior to 2005 that searched a single database, or did not undertake a risk of bias assessment.

Diagnostic tests considered in the OvR included clinical examinations, radiologic tests (i.e., X-ray, mylography, electro-diagnostics, CT, MRI, provocation discography), specialized clinical tests (i.e., ROM, muscle endurance and strength, palpation tests, provocation tests for nerve tissue, functional tests, non-organic signs), injections, laboratory tests, blood tests; no citation was excluded based on the types of diagnostic tests. Articles that attempted to evaluate different systems of NP classification systems, consensus processes or models, definitions, terminologies, or frameworks aimed at categorization of NP diagnoses or disorders were also eligible. Exclusions: None for diagnostic tests or classification systems of the neck.

For reviews evaluating the evidence for prognostic factors, any prognostic factor or predictor (i.e., accident parameters, biological factors, psychological factors, behavioral factors, symptoms or interference factors, medico-legal context factors, other social and demographic factors) of outcomes were eligible; studies evaluating primary risk factors for the onset of neck-related problems were excluded. Articles were not restricted by the type of outcome associated with the prognostic factor in patients with NP. Where described, outcomes were grouped according to those types that could be included within the general International Classification of Function (ICF) domains of body structure and function, activity limitation, restricted participation, and environmental factors. Exclusions: None for prognostic factors; SR of risk factors for developing NP were excluded.

For reviews evaluating outcome measures, publications that evaluated at least one of the psychometric/clinimetric properties (reliability, validity, responsiveness, cross-cultural validation, floor-ceiling effect, interpretability) of self-reported or observed outcomes were eligible. There were no restrictions on the type of outcome measure evaluated within NP populations. Exclusions: None for type of outcome measure.

Types of Publications and Language Restrictions for Reviews and CPG

Eligible publications included systematic or narrative reviews [7], other OvR, CPG or consensus statements or commentaries associated with NP. Treatment, harms, and prognosis areas were limited to SR only; diagnosis and outcomes allowed narrative reviews and commentaries related to the area being evaluated. SR are characterized by comprehensive methods to identify and synthesize all the literature on a given topic. A publication was considered to be a narrative review if it labeled itself as a review but did not clearly identify the methods for selecting relevant studies or synthesizing the evidence. All CPG and consensus statements that included recommendations or algorithms for the management of NP were eligible. Note that CPGs are defined as systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances; they can be developed by local, regional, national or international groups or affiliated governmental organizations. Consensus statements are similar; but reflect a different methodology for deriving recommendations. Algorithms are variable in how they present guidance and may reflect recommended pathways for treatment or diagnosis.

Search Methods for Identification of Studies

A research librarian searched the following computerized bibliographic databases of the medical, chiropractic, and allied health literature, without language restrictions from January 2000 to March 2012: MEDLINE, EMBASE, CINAHL, ILC, and CENTRAL, and LILACS. References within eligible articles were screened for any additional references, as well as relevant publications from personal files of the investigative team. **APPENDIX** shows the search terms used for the OvR in MEDLINE.

Data Collection and Analysis

Selection of Studies

Screening of titles and abstracts focused on identifying the following: systematic or narrative reviews, other OvRs, CPGs or consensus statements associated with NP, in any of the five clinical areas (with exclusions as described above). Two screeners independently evaluated the titles and abstracts or full text articles from the computer searches using pre-piloted forms placed within Distiller software (copyright 2009). Disagreements were resolved through consensus in both stages of selection. If consensus could not be reached, a third rater (lead investigator) adjudicated the final rating. If the article or citation posting was in a language other than English, one investigator and a translator with a medical background conducted the study selection and extraction.

Data Extraction and Synthesis

A single reviewer extracted primary fields related to the characteristics, main findings, risk of bias assessment, and strength of evidence evaluation in SR and quality of CPGs. Each of the primary authors then independently verified the extracted data and inconsistencies were resolved by consensus. The specific fields extracted are detailed in the individual summary papers (see this issue for some OvR in treatment, prognosis and outcomes). Generally, characteristics of the eligible reviews (and characteristics of the studies evaluated within the reviews) were extracted and included types of NP, types of interventions, diagnostic tests, prognostic factors, and outcomes. Qualitative syntheses were undertaken in each of the OvR for the five clinical areas of NP management; the potential for quantitative synthesis was evaluated separately for each area. Where possible, summary tables of the primary findings of each review were assembled and presented. Synthesis of SR and narrative reviews were summarized separately.

Assessment of Risk of Bias in Included SR and CPG

Two screeners applied the AMSTAR risk of bias tool for SR. This checklist has been shown to have good face and content validity for assessing methodological quality [8, 9]. The AMSTAR evaluates 11 criteria within SR that have the potential to introduce systematic bias. Some of the items considered within the AMSTAR tool include comprehensiveness of the search strategy (including grey literature), the degree to which characteristics of the primary studies are adequately presented, the method for assessing scientific quality of eligible studies, the appropriateness of methods used to synthesize studies, identification of publication bias, and conflict of interest.

CPG were evaluated using the AGREE II instrument [10] by two screeners. The AGREE II evaluates six domains that assess the process and rigour of CPG development. These included the domains of scope and purpose, stakeholder involvement, rigour of development, clarity of presentation, applicability, and editorial independence. Scores were summarized for each of these domains. The AGREE II has been shown to be a reliable and valid instrument to critically appraise guideline development [11].

Qualitative Analysis of the Strength of Evidence

Currently, no methods exist to evaluate the strength of evidence (SOE) for findings across different SRs. Approaches to assess SOE, such as GRADE, are currently used with primary studies from a single systematic review [12-18]. In general, GRADE criteria used to judge the primary studies within a SR include the study design, risk of bias; imprecision (sample size); inconsistency (direction and magnitude of results); indirectness (applicability to patient populations with NP); and reporting bias.

As such, our team developed an approach using the GRADE but adapted for SR (rather than primary studies) evaluating the efficacy of treatments and diagnostic tests. Where available the rating of SOE within each review for specific interventions were extracted; when a SOE rating was not reported, an attempt was made by investigators to rate this evidence using the GRADE approach. If more than one SR was eligible for one type of intervention (e.g. three SR on exercise and NP)) and only one provided a rating of SOE using the GRADE approach, then that SOE reported was the one that was extracted and reported (for all three SR in this example). When SR reported SOE using different systems, these were compared to those SR that did use the GRADE approach; any similarities or differences between SOE ratings in these reviews were compared and a final judgement was made for the specific intervention. When there were two or more SR evaluating the same intervention and that provided GRADE SOE ratings, then all ratings were extracted and reported. Attempts were made to reach a consensus with respect to a final rating for the SOE for each intervention. A qualitative summary of the SOE using the same domains (risk of bias, inconsistency, imprecision, indirectness, and publication bias) is presented for evaluating the overall quality of OvR for prognostic or outcome measurement studies; currently no guidelines are provided for prognostic or outcome type studies using the GRADE system. Magnitude and consistency of synthesized results across different SR teams provided the final SOE statement in these areas. For the CPG, the proportion achieving greater than 50% in each of the six domains within the AGREE II is used as the threshold to characterize the overall quality of the guidelines.

RESULTS

Fig. (**1**) shows the flow of eligible articles for reviews evaluating all the clinical areas for NP management. When combined across the different areas of NP management, a total of 10,059 articles were found and from these, 7579 were duplicates (75%). Articles from grey literature sources and personal files (n=101) are also shown in Fig. (**1**) and these were relevant primarily in the OvR of treatment and CPG. From 2480 initial articles (after duplicates were removed), 512 were screened at full text. Fig. (**1**) also shows the distribution of eligible articles with respect to the areas of management of NP; there was some overlap between clinical areas (for example, there were 12 articles that were eligible for both diagnosis and outcomes). This overlap is a reflection of the content of the publications containing information relevant to both areas.

Table **1** shows the number of duplicates amongst the different bibliographic databases and the yield for the different clinical areas that were searched. From these, the search for reviews of outcome measures yielded the smallest proportion of overlap when searching across different databases and the greatest overlap was in prognosis (54%).

Relative Distribution of Literature on NP

Fig. (**1**) shows that the majority of eligible articles were for the clinical area of treatment (n=117) and the smallest number for harms and prognosis (n=16 each). Diagnostic tests also yielded a relatively large number of articles of which the majority were SR (44 from 54). When considering the reasons for exclusion at full text, the majority of excluded articles were treatment reviews published prior to 2005 (n=103). Similarly, a large number of articles that evaluated chronic pain management for musculoskeletal disorders did not have sufficient (greater than 50% studies on NP) or stratified data for NP and were excluded (n=73). Fig. (**1**) shows other reasons for exclusion at full text. It was anticipated that some areas of NP management would not be well indexed or have few SR. As such, narrative reviews and commentaries were eligible for diagnosis/classification, prognosis and outcomes. Although a large number of articles for diagnosis and classification (n=109) were eligible given the intentionally broad criteria during screening, the greatest proportion of articles were narrative reviews or commentaries on classification of neck disorders (n=55). However, a preliminary analysis of the articles for classification revealed a disparate set of publications that were not primarily focused on summaries or evidence for disease taxonomy. There was often overlap with other clinical areas and very little data to extract for classification. As such, these 55 articles are now excluded from this OvR (see Fig. **1**). Table **1** shows that the OvR of diagnosis and outcomes included a small proportion of narrative reviews in their final inclusion that provided summary information on a specific diagnostic test or outcome measures.

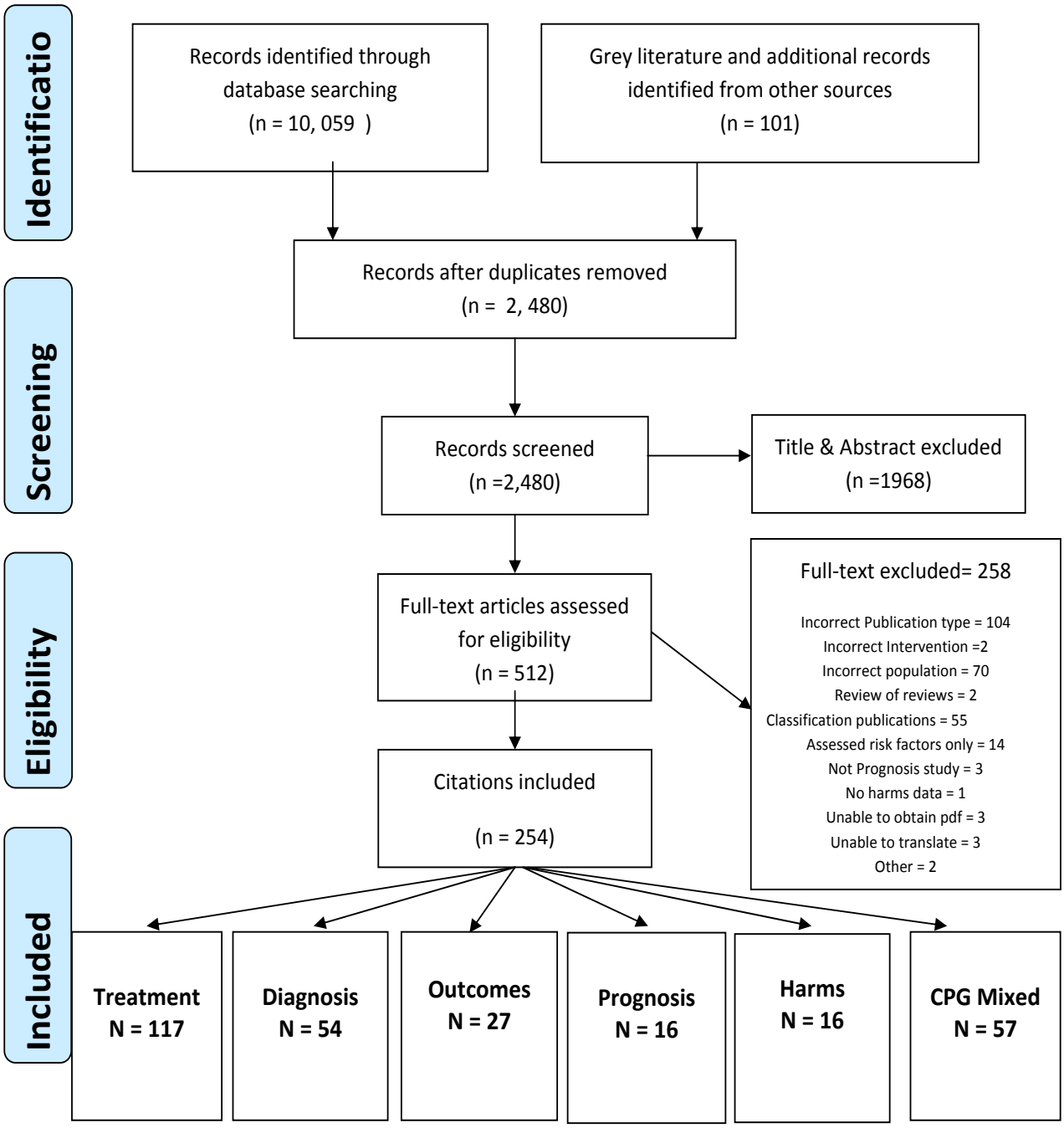

Fig. (1). Flow Diagram of citations screened and eligible for this review of reviews.

Table 1. Yield of Citations Across the Different Areas of NP Management

	Treatment or Harms (%)	Diagnosis and Classification	Prognosis	Outcome Measures
Citation Yield	5056	1991	2518	490
Duplicates across searched databases	2048 (41%)	836 (42%)	1371 (54%)	95 (19%)
Interim Yield	Tx-SR =113 Hx-SR = 16	109 (SR = 44/ NRO = 65)	SR = 16	27 (SR = 13/ NRO = 14)
Final Yield	Tx-SR = 113 Hx-SR = 16	SR /NRO = 54	SR = 16	SR/NRO = 27

Tx = treatment; Hx = Harms; SR = systematic review; NRO = narrative reviews

Risk of Bias

Figs. (**2**, **3**) compare the criteria for AMSTAR across the areas of NP management. AMSTAR scores for each individual area of NP management are shown in web **APPENDIX** (Figs. **1-5**). There were three criteria where generally all clinical areas were at high risk of bias (Fig. **2**). A specific conflict of interest statement (AMSTAR Q11) was predominately absent (88% to 100% of SRs) in all areas of NP management.

Assessment of publication bias (AMSTAR Q10) was also absent in most areas of NP management (96% to 100% of reviews). Similarly, the majority of SR did not report the list of excluded studies (AMSTAR Q5) (61% to 94%) or used the publication status as an inclusion criteria (AMSTAR Q4) (43% to 92%). With the exception of treatment articles (56%), SR in the other clinical areas employed "duplicate" selection and data extraction of eligible studies (AMSTAR Q2) infrequently (25% to 46%).

In contrast, all areas of NP management scored well on a priori specification of the protocol (AMSTAR Q1) (88% to 100%) (Fig. **3**). With the exception of the areas of prognosis (57%), all other areas generally undertook a comprehensive search (AMSTAR Q3) (80% to 100%). Generally, assessment of the risk of bias (AMSTAR Q7) and appropriate use of this in the formulation of conclusions (AMSTAR Q8) was achieved fairly well in the treatment and prognosis SR (89/85% to 93/86%); this was less consistently undertaken in the areas of diagnosis, outcomes, or harms (54% to 57%). Note that the AMSTAR does not assess the merits of the various methods used to assess risk of bias of the primary studies included in each SR. With the exception of diagnostic and outcome reviews (46% to 69%), the SR in the other three clinical areas adequately described the characteristics of included studies (AMSTAR Q6) (81% to 86%).

Meta-analyses were rarely undertaken in (AMSTAR Q9) the SR and as such most areas (68% to 100%) received a 'not applicable' rating for methods used to combine studies (even in SR of treatment). The reporting within the reviews was sufficiently ambiguous that it could not be determined if the studies undertook duplicate study selection and extraction (AMSTAR Q2) from 10% to 12% for treatment and harms and 20% to 23 % in the other areas.

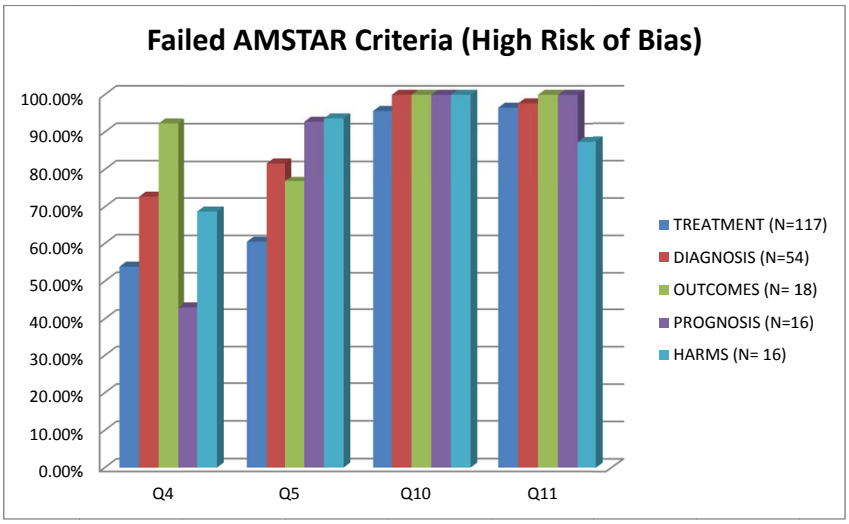

Fig. (2). Areas of neck pain management and AMSTAR scores indicating high risk of bias for these domains.

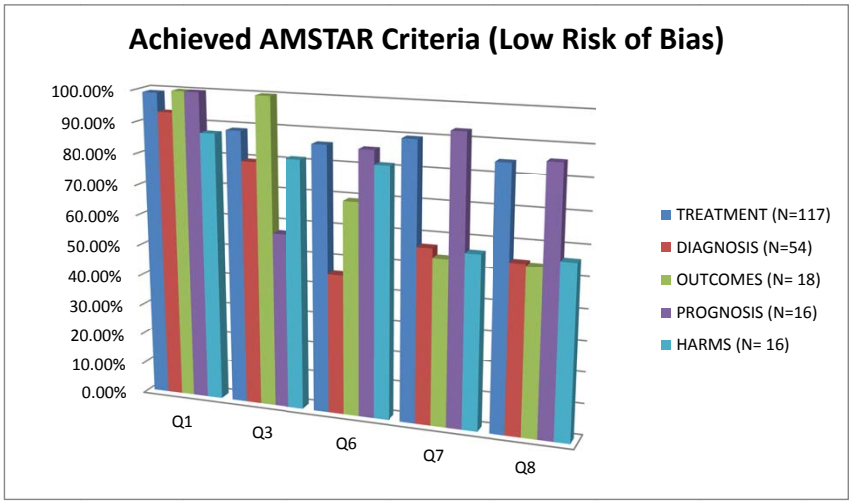

Fig. (3). Areas of neck pain management and AMSTAR scores indicating low risk of bias for these domains.

DISCUSSION

This paper describes the methods used to undertake a comprehensive OvR to evaluate the evidence in the five key areas in the management of NP. The motivation in undertaking an OvR was to efficiently summarize the state of the evidence across a large literature base that addresses the major aspects of management of NP. Since an OvR serves to provide a high-level or panorama view of the literature in a specific area [19] comparisons across clinical areas is ideal; as such, the relative amount of literature and overlap across clinical areas, as well as any differences in AMSTAR scores risk of bias were evaluated in this OvR. The specific findings of the OvR within each clinical area of NP management are presented elsewhere.

There are some potential limitations to the approach taken in the methods employed in this OvR. When considering threats to validity of OvR, several issues have been identified in the literature. There is some potential risk for selective outcome reporting even in OvR and therefore a priori specification of the outcomes within the OvR would be recommended [20]. The methodologies in this OvR did not restrict SR based on the outcomes reported in any of the five clinical areas; however, within each specific clinical area, there was some variation in which outcomes were reported; this challenge is similar to differences across primary studies summarized within SR. Experience with OvR including only Cochrane SR have suggested greater attention to reporting findings based on the outcome (rather than the intervention) to facilitate interpretation [20]. Although there are no specific recommendations in the literature for assessing SR of diagnosis, harms, prognosis, or outcomes, this suggestion may be equally applicable to these areas.

Another issue that may affect validity of the OvR findings concerns the recency of the eligible SR; if reviews are not up to date, then the validity of the results are compromised. Our search commenced in 2000, and captures approximately a 12 year interval; an interval greater than five years was selected, as it was anticipated that some types of interventions, diagnostic tests, or CPG would not have had a summary evaluation in this timeframe. For the articles evaluating treatment interventions, the overwhelming number of SR available necessitated re-evaluation of the time interval and as such the eligible articles were restricted to the last five years only. This approach would be consistent with those proposed by Whitlock *et al.* (2008)[5] who recommend the selection of SR that are the most chronologically recent and at the lowest risk of bias. The prognostic SRs were limited to those published in the previous 10 years (back to 2002), since many of the primary papers were included in multiple SRs, and inclusion of the previous 10 years provided ample saturation of this literature. In order to include all relevant interventions, clinical tests, harms, or outcomes used to manage NP, SR with publication dates prior to 2005 were included. It can be reasoned that less current evidence (older than 5 years) is preferable to presenting none; the lack of updates in any specific clinical area serves as good indicator of where future research efforts could be directed.

Other general issues when undertaking OvR concern potential difficulties in combining findings from different meta-analyses estimated in the original SR; however, none of the specific clinical area OvR undertook computations of an overall summary estimate. Although methods to summarize evidence across SR in OvR have been established [1,5], there are variations in approaches used that have been identified in the literature [21, 22]. Most notably, some OvR have not assessed risk of bias in the original SR, or addressed discrepancies between SR. Methods of addressing discordance amongst reviews were not specified a priori, but in general explanation of differences in direction and magnitude of effects, and differences in methodologies within the SR were noted as previously specified [23]. The methods employed in this OvR have been explicitly described and eligible SR have been assessed for risk of bias and reporting of findings attempted to comply with reporting standards for systematic reviews, albeit not developed for OvR [24].

Additionally, the development of our SOE approach required some a posteriori consensus for assigning a level of evidence when one was not provided or provided in a system other than GRADE. An adaptation of the GRADE system was undertaken where the same domains (risk of bias, inconsistency, indirectness, imprecision, and publication bias) were considered in the general sense for SR. OvR conducted within the Cochrane Collaboration recommend the reporting of methods used to determine SOE within eligible SR [20]. However this task is relatively simple in Cochrane reviews as standardization of reporting SOE are based on the GRADE. The limitations to the approach for extracting and summarizing SOE used in this OvR are not yet known, but have face validity.

Yield and Overlap Among the Areas of NP

Our yield of articles across the different clinical areas shows a significant gap with respect to areas of inquiry on the management of NP. This search of reviews showed a primary focus of all publications to be on treatment of neck pain and to a lesser extent for diagnostic tests; the numerous treatments and clinical tests may in part account for this. The majority of CPG also focused on treatment recommendations, with a lesser emphasis on diagnostic tests. In contrast, relatively few SR evaluated the potential for harms, prognostic factors, or the use or properties of outcome measures. Previous research [25, 26]. would suggest a different distribution of literature across the clinical areas, with studies on risk factors and treatment for neck pain being most frequent; the area most infrequently assessed in this previous OvR was articles assessing economic costs. This previous research included predominately primary studies rather than SR. Owing to our purpose to use the results of these OvRs to inform clinical practice in all areas of NP management, we were not concerned with SRs of primary risk factors for the onset of *new* neck pain, rather our focus was on management, measurement and prognosis (secondary risk) of existing neck pain. Harms, outcome measures, and classification studies are novel additions of the current series of OvRs, having not been evaluated in the previous OvR.

When considering duplication of articles across different bibliographic databases, it was surprising that the SR on outcome measures yielded the smallest proportion of overlap; this suggests that searching multiple databases are important for this clinical area of NP management. Additionally, there was a fair amount of overlap in articles across the different clinical areas. This may reflect a trend in assembling more comprehensive summaries addressing multiple areas, especially combined guidelines and best evidence syntheses.

The screening yielded a large number of articles for diagnosis and classification (n=109), and the greatest proportion were narrative reviews on classification of single or multiple NP disorders or syndromes (n=55). Our criteria for screening in classification-related articles were very broad and may have been a factor in this large number. However, a preliminary analysis of the articles for classification revealed a disparate set of publications that were not primarily focused on summaries or evidence for disease taxonomy. There was overlap with other clinical areas and very little data to extract for classification. This suggests that our search terms, which were limited to systematic and narrative reviews, were not sufficiently comprehensive. A review of primary studies may be required to adequately address the evidence for classification of NP.

Diagnostic Tests versus Outcome Measures

There were a number of methodological issues that ensued while undertaking this OvR. One issue that required some a posteriori consensus concerned the definition of an outcome measure versus a clinical or diagnostic test. A traditional conceptualization of a diagnostic or clinical test was one that would assist in confirming the presence or absence of disease. However, recent conceptualizations of clinical tests have been expanded to broaden this scope, such that any process that is used to inform or change patient management can be regarded as a clinical test [27]. Clinical tests can be used to monitor disease changes, assess prognosis, and screen for presence of disease. Parallel to this is more recent definitions of an "outcome measure", which defines the variety of outcomes as ones that provide clinical "endpoints" (that may include morbidity, health or functional status, mortality, and even laboratory values) [5]. The distinction between a diagnostic or clinical test and an outcome measure appears to be less clear. This is also evident in clinical practice where a clinical test can be used several times for different purposes. As an illustrative example, one might consider range of motion testing, where this is typically used to assess current mobility status at initial assessment (diagnose movement problems), it can also be used to monitor the impact of treatment, as part of a final assessment to indicate the end of treatment (outcome), or as a prognostic factor following traumatic neck injury. Similarly, the Neck Disability Index can be used to assess severity of functional limitations and be used by a clinician to stratify patients into prognostic groups or to inform selection of treatment options (based on these groupings), but it can also be used to assess the outcome of the treatment. Thus, the same clinical test/outcome measure can be used for assessment (or diagnostic) purposes, as well as for prognosis, classification, or measurement of treatment outcome (outcome measure).

From a practical perspective distinguishing where the eligible reviews should be categorized with respect to placement within the diagnostic, prognostic or outcome groups in some cases proved challenging. One attempt was to distinguish the aim of the review by the types of the analyses undertaken, but this too proved problematic. Often the eligible citation would evaluate a clinical test using traditional accuracy performance measures (i.e., sensitivity and specificity) and also clinometric characteristics (reliability, validity, responsiveness). Our attempt to resolve this classification dilemma was to include the articles in several of the clinical areas in consultation with the expert panel, many of whom had representation on the included SRs. Nonetheless, this overlap may serve to inflate the relative proportion of literature in these three areas and report findings in duplicate across the different clinical areas.

CONCLUSIONS

We have provided a description of the methods used to undertake a comprehensive review of published SR across all areas of NP; this OvR captured relevant reviews involving treatment, harms, diagnosis, prognosis and outcomes for the management of NP. The challenges of selecting and synthesizing eligible articles in an OvR required customized solutions across different areas of clinical focus. Comparisons across the different areas of NP show that the greatest number of SR evaluated treatment interventions and the fewest outcome measures. Some differences in the risk of bias with individual SR across the different areas of NP were also noted; generally there were consistent problems with reporting potential conflict of interest, assessing publication bias, specification of excluded studies list, and inclusion of grey literature. Future SR could address the noted risk of bias deficiencies and the relative imbalances of knowledge for the different areas of NP management.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by Canadian Institutes of Health Research (CIHR) grant(s) FRN: KRS-102084.

The ICON authors that provided direction of the project and reviewed the findings/manuscript include (in alphabetical order): Gert Bronfort, Norm Buckley, Lisa Carlesso, Linda Carroll, Pierre Côté, Jeanette Ezzo, Paulo Ferreira, Tim Flynn, Charlie Goldsmith, Anita Gross, Ted Haines, Jan Hartvigsen, Wayne Hing, Gwendolen Jull, Faith Kaplan, Ron Kaplan, Helge Kasch, Justin Kenardy, Per Kjær, Janet Lowcock, Joy MacDermid, Jordan Miller, Margareta Nordin, Paul Peloso, Jan Pool, Duncan Reid, Sidney Rubinstein, P. Lina Santaguida, Anne Söderlund, Natalie Spearing, Michele Sterling, Grace Szeto, Robert Teasell, Arianne Verhagen, David M. Walton, Marc White.

APPENDIX

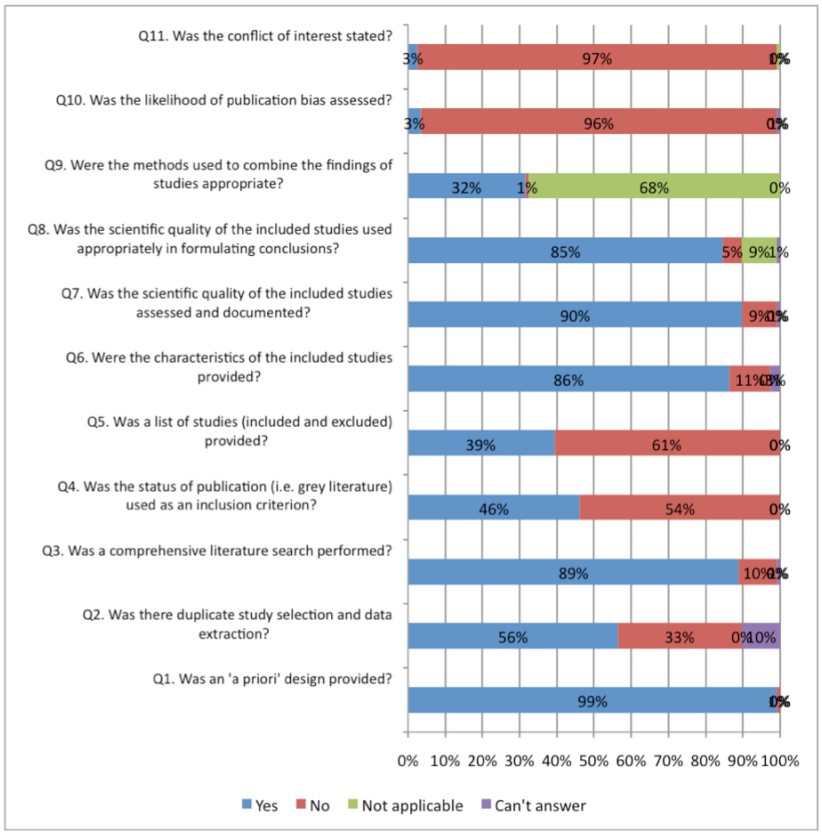

Fig. (1). AMSTAR scores for reviews evaluating treatment of NP.

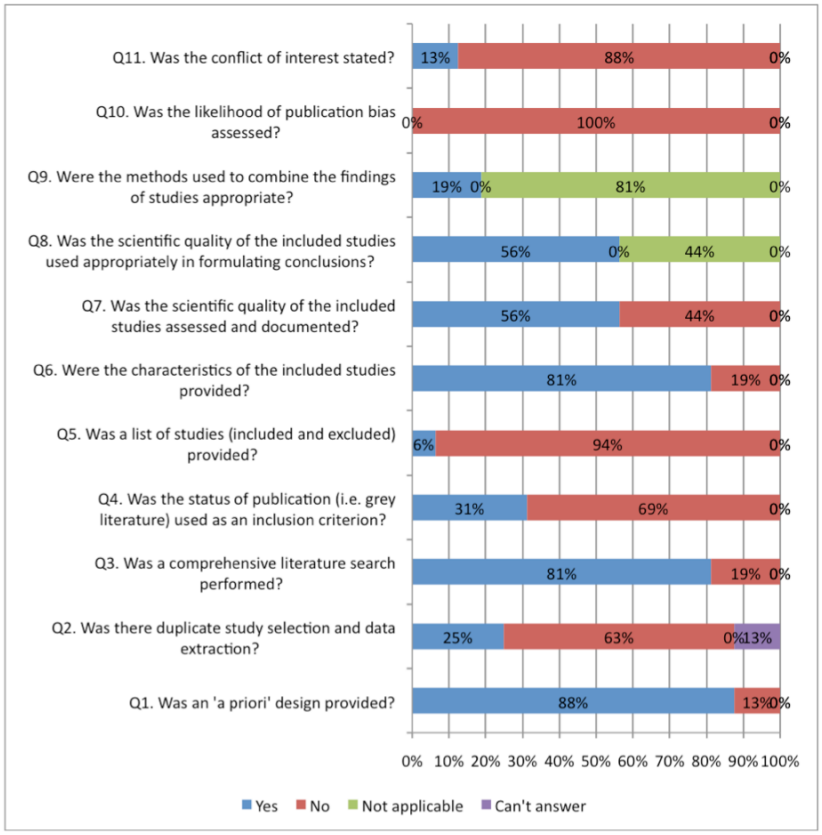

Fig. (2). AMSTAR scores for reviews evaluating harms of interventions used in NP.

(Appendix) contd…..

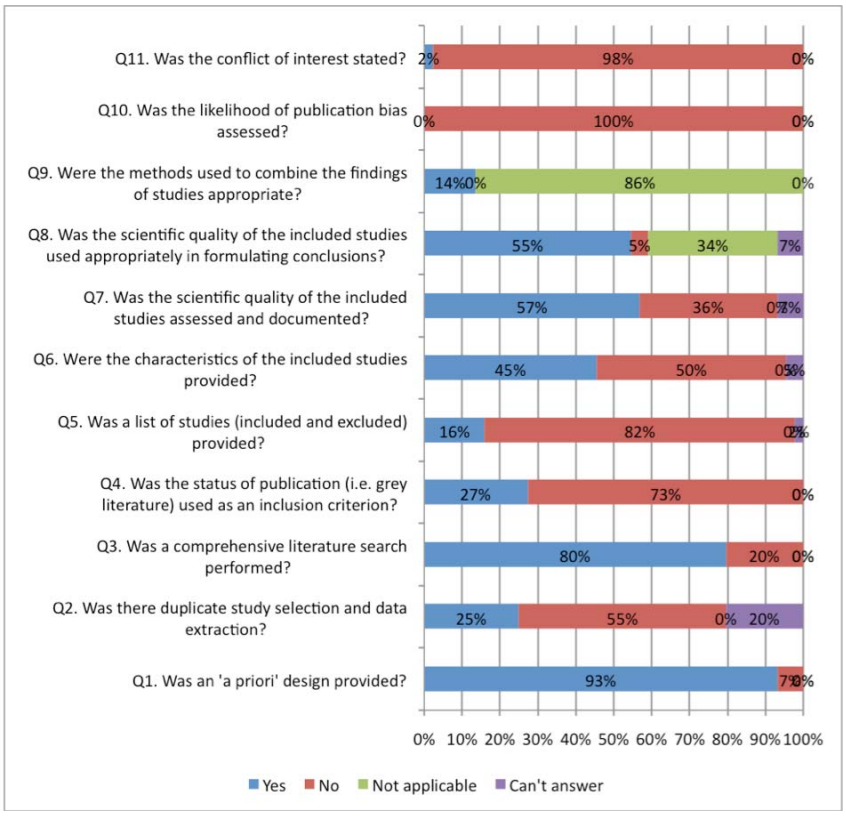

Fig. (3). AMSTAR scores for reviews evaluating diagnostic tests used to assess NP.

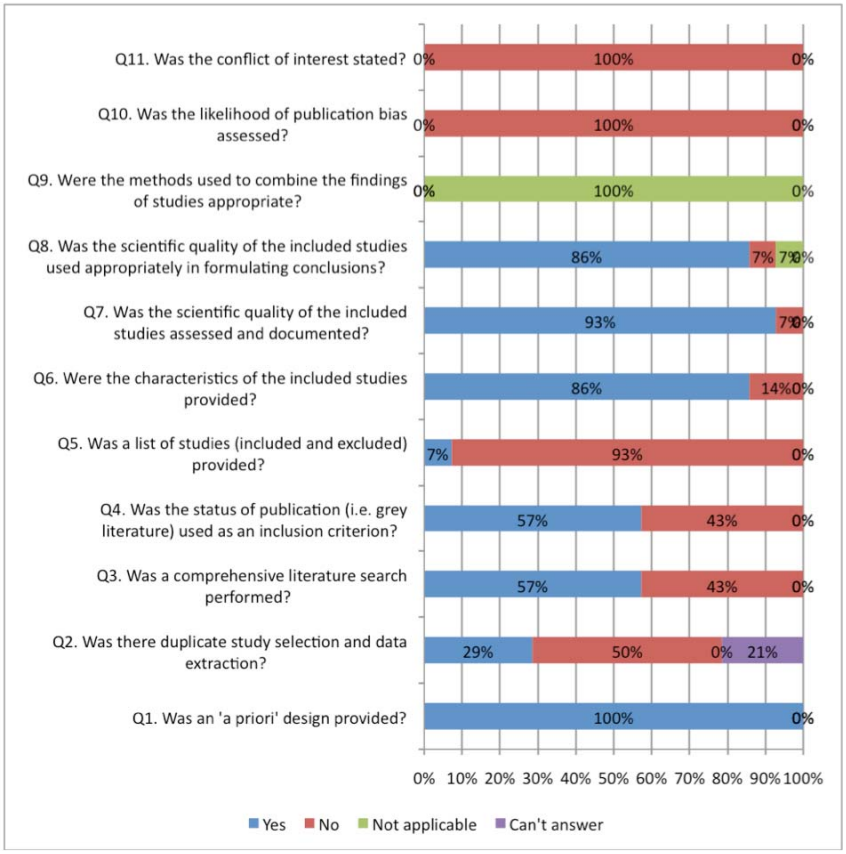

Fig. (4). AMSTAR scores for reviews evaluating prognostic indicators of NP.

(Appendix) contd…..

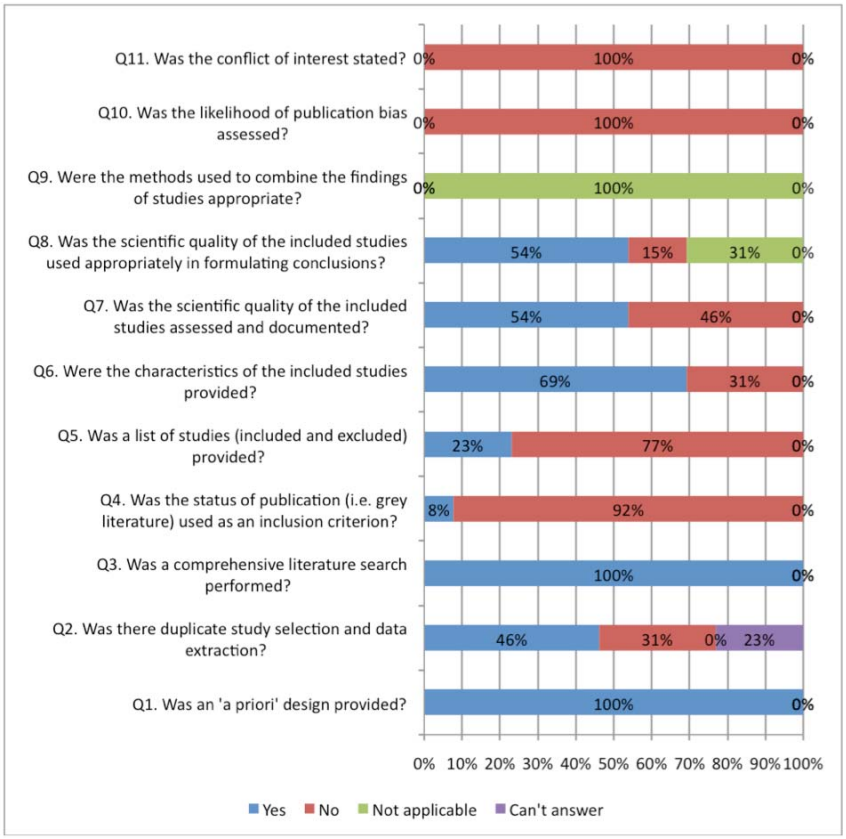

Fig. (5). AMSTAR scores for reviews evaluating outcome measures used in NP.

REFERENCES

[1] Becker LA, Oxman AD. Chapter 22: Overviews of reviews. In: Higgins JPT, Green S, Eds. Cochrane Handbook for Systematic Reviews of Interventions. 5.0.1 ed. The Cochrane Collaboration 2011.

[2] John PA, Ioannidis MD. Integration of evidence from multiple meta-analyses: A primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. CMAJ 2009; 181(8): 488-93.

[3] White CM, Ip S, McPheeters M. Using existing systematic reviews to replace de novo processes in conducting comparative effectiveness reviews. In: Agency for Healthcare Research and Quality, editor. Methods Guide for Comparative Effectiveness Reviews. Ed. Rockville: Agency for Healthcare Research and Quality 2009.

[4] Carroll LJ, Hogg-Johnson S, van d, V, *et al.* Course and prognostic factors for neck pain in the general population: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. J Manipulative Physiol Ther 2009; 32(2 Suppl): S87-S96.

[5] Whitlock EP, Lin JS, Chou R, Shekelle P, Robinson KA. Using existing systematic reviews in complex systematic reviews. Ann Intern Med 2008; 148(10): 776-82.

[6] Gross A, Miller J, D'Sylva J, *et al.* Manipulation or mobilisation for neck pain. Cochrane Database Syst Rev 2010; (1): CD004249.

[7] Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. Ann Intern Med 1997; 126(5): 376-80.

[8] Shea BJ, Grimshaw JM, Wells GA, *et al.* Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007; 7: 10.

[9] Shea BJ, Bouter LM, Peterson J, *et al.* External validation of a measurement tool to assess systematic reviews (AMSTAR). PLoS ONE 2007; 2(12): e1350.

[10] AGREE Next Steps Consortium. The AGREE II Instrument. http://www.agreetrust.org 2009. Available from: http://www.agreetrust.org

[11] Brouwers MC, Kho ME, Browman GP, *et al.* AGREE II: advancing guideline development, reporting and evaluation in health care. J Clin Epidemiol 2010; 63(12): 1308-11.

[12] Guyatt GH, Oxman AD, Kunz R, *et al.* GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol 2011; 64(12): 1294-302.

[13] Guyatt GH, Oxman AD, Kunz R, *et al.* GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol 2011; 64(12): 1283-93.

[14] Guyatt GH, Oxman AD, Montori V, *et al.* GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol 2011; 64(12): 1277-82.

[15] Guyatt GH, Oxman AD, Kunz R, *et al.* GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol 2011; 64(12): 1303-10.

[16] Guyatt GH, Oxman AD, Sultan S, *et al.* GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011; 64(12): 1311-6.

[17] Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011; 64(4): 380-2.

[18] Guyatt GH, Oxman AD, Kunz R, *et al.* GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011; 64(4): 395-400.

[19] Ioannidis JP. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. CMAJ 2009; 181(8): 488-93.

[20] Thomson D, Russell K, Becker L, Klassen T, Hartling L. The evolution of a new publication type: Steps and challenges of producing overviews of reviews. Res Synth Methods 2011;1(3-4):198-211.

[21] Woodman J, Thomas J, Dickson K. How explicable are differences between reviews that appear to address a similar research question?

A review of reviews of physical activity interventions. Syst Rev 2012; 1(1): 37.

[22] Pieper D, Buechter R, Jerinic P, Eikermann M. Overviews of reviews often have limited rigor: a systematic review. J Clin Epidemiol 2012; 65(12): 1267-73.

[23] Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. CMAJ 1997; 156(10): 1411-6.

[24] Liberati A, Altman DG, Tetzlaff J, *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009; 6(7): e1000100.

[25] Cassidy JD, Cote P. Is it time for a population health approach to neck pain?. [Review] [40 refs]. J Manipulative Physiol Ther 2008; 31(6): 442-6.

[26] Carroll LJ, Cassidy JD, Peloso PM, *et al*. Methods for the best evidence synthesis on neck pain and its associated disorders: results of the bone and joint decade 2000-2010 task force on neck pain and its associated disorders. J Manipulative Physiol Ther 2009; 32(Suppl 2): S39-45.

[27] Centre for Reviews and Dissemination (CRD). Systematic Reviews: CRD's guidance for undertaking reviews in health care. NY: University of York 2009.

Received: December 11, 2012 Revised: March 19, 2013 Accepted: March 19, 2013