Musculoskeletal Science and Practice 42 (2019) 125–133

Contents lists available at ScienceDirect

Musculoskeletal Science and Practice

journal homepage: www.elsevier.com/locate/msksp

Systematic Review

What is the diagnostic accuracy of red flags related to cauda equina syndrome (CES), when compared to Magnetic Resonance Imaging (MRI)? A systematic review

Nathalie Dionne[a], Abiodun Adefolarin[a], Dena Kunzelman[a], Nitin Trehan[a], Laura Finucane[b], Lenerdene Levesque[a], David M. Walton[a], Jackie Sadi[a,*]

[a] Western University, School of Physical Therapy, Rm 2305 Elborn College, London, ON, N6G 1H1, Canada
[b] Sussex MSK Partnership, 177 Preston Road, Brighton, BN1 6AG, UK

ARTICLE INFO

Keywords:
Cauda equina syndrome
Low back pain
Signs and symptoms
Red flags
Accuracy
MRI

ABSTRACT

Objective: To review and statistically pool available evidence on the diagnostic accuracy of red flags to clinically identify MRI confirmed Cauda Equina Syndrome (CES).
Study design: Systematic review.
Data sources: Embase, Scopus, Ovid Medline, Ovid Healthstar, Amed and CINAHL from inception to January 30, 2018 and a grey literature search.
Inclusion criteria: Primary diagnostic studies, published in English; comparing red flags for CES; to Magnetic Resonance Imaging (MRI) as reference standard; in humans; older than 18 years.
Methods: Data extraction, assessment of study quality using a modified QUADAS-2 tool and the use of GRADE to synthesize the results for each test was performed by three independent assessors. Diagnostic accuracy statistics applied to the identified data andpooled analysis performed using Meta-DiSc, version 1.4. Moderator analyses planned for pooled results.
Results: Seven studies (total N = 569 participants) were included. Potential signs or symptoms of CES were compared to MRI findings. Diagnostic data could be pooled for reduced anal tone, leg pain, back pain, saddle anaesthesia, urinary retention, urinary incontinence and bowel incontinence from six of seven studies. The pooled sensitivity for the signs and symptoms ranged from 0.19 (95% CI 0.09 to 0.33) to 0.43 (95% CI 0.30 to 0.56) while the pooled specificity ranged from 0.62 (95% CI 0.59 to 0.73) to 0.88 (95% CI 0.85 to 0.92).
Conclusion: Red flags used to identify potential CES appear to be more specific than sensitive. As such, when these are present, they should be considered justification for prompt diagnostic workup.

1. Introduction

The cauda equina consists of a group of lumbar and sacral nerve roots that originate from the conus medullaris at the base of the spinal cord. It begins just below the first lumbar vertebra and provides sensory as well as motor innervation to the lower extremities, bladder and bowel and plays a role in sexual function. Cauda Equina Syndrome (CES) is a serious pathology that occurs when there is compression of the cauda equina which can lead to permanent loss of bowel and bladder control, loss of sensation in the perineal and saddle region and of the lower extremities as well as weakness of the lower quadrant. The incidence of CES within the United Kingdom was estimated at 1.9 per 100 000 (Woodfield et al., 2018). A herniated intervertebral disk in the lumbar spine is the most common cause of this peripheral neuropathic condition, with 2% of all herniated disks resulting in CES (Bydon et al., 2016). Without rapid identification and management, CES can lead to life-changing, irreversible disability.

Clinicians are advised to assess whether serious underlying pathology exists in patients with low back pain by evaluating the presence of key clinical signs and symptoms, referred to as “red flags” for CES, during the subjective history and physical examination. These are described as signs and symptoms thought to indicate a higher risk of serious pathology and warrant referral for further diagnostic testing (Henschke et al., 2013). Red flags associated with CES are loss of sensation in the saddle and perineal region, urinary retention/overflow incontinence, fecal incontinence, reduced anal sphincter tone, sciatica,

* Corresponding author. Western University, School of Physical Therapy, Rm 2305 Elborn College, London, ON, N6G 1H1, Canada.
E-mail address: jsadi2@uwo.ca (J. Sadi).

https://doi.org/10.1016/j.msksp.2019.05.004
Received 20 December 2018; Received in revised form 26 April 2019; Accepted 11 May 2019

hyporeflexia and motor/sensory loss in the lower limbs (Fraser et al., 2009; Germon et al., 2015; Lavy et al., 2009; Verhagen et al., 2016). Most guidelines recommend that patients presenting with any of these key clinical signs and symptoms should be referred for urgent Magnetic Resonance Imaging (MRI) in order to confirm the presence of CES (Bell et al., 2007, Fairbank et al., 2011).

Much of the evidence informing the utility of the key clinical signs and symptoms in screening for CES has come from anecdote or research studies of limited sample size. Knowledge syntheses such as systematic reviews and meta-analyses are appropriate ways of improving point estimates and statistical power in a field especially where many small studies exist.

This systematic review provides part of the basis of an International research study aiming to build a consensus on which red flags might be useful in the identification of serious spinal pathology and how they should be used in the clinical setting. The need for further research to provide data on the diagnostic accuracy of red flags was identified following a consultation of the member organizations of the International Federation of Orthopaedic Manipulative Physiotherapists (IFOMPT) a sub group of World Confederation of Physical Therapy (WCPT) in 2016. The four key areas where systematic reviews are urgently required were CES, infection, fracture and malignancy. The purpose of this study was therefore to systematically identify prior research exploring the diagnostic accuracy of key clinical signs and symptoms associated with CES red flags in adults with low back pain, and to compare the utility of those findings using MRI confirmed compression as a reference standard. Where adequate data existed, this study also planned to conduct statistical pooling to improve estimates of the true diagnostic accuracy.

2. Methods

2.1. Data sources

Embase, Scopus, Ovid Medline, Ovid Healthstar, Amed and CINAHL were searched from inception to January 30, 2018. The searches used combinations of terms related to CES signs, symptoms, physical examination, red flags, reliability, accuracy and MRI. The full search strategy is provided in Appendix A. A grey literature search was conducted in non-database sources (google scholar, physiopedia), article reference lists and government clinical trial registries.

2.2. Study selection

Primary diagnostic studies were considered if they examined the results of physical examination and/or subjective history for signs and symptoms related to CES. Since MRI is widely accepted as the reference standard for CES, only studies that compared the clinical findings with MRI were included. Studies presenting sufficient data to allow extraction and pooling of estimates of diagnostic accuracy were included. We considered studies in English, involving human adults ($\geq$18 years of age) who presented with suspected CES from an insidious onset or herniated disc prolapse. Studies on CES caused by a disease process were excluded however (infection, malignancy, trauma, autoimmune, and congenital disorders). Two authors independently reviewed titles and abstracts, and final selection was based on a review of full articles. A Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)(Moher et al., 2009) flow diagram was generated depicting the flow of information, during the different phases of this study (Fig. 1, Appendix A).

2.3. Quality assessment

For the purpose of this study, the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool (Whiting et al., 2011) was piloted for improved clarity and consistent grading given the context of this study (Appendix B). Two of the authors independently examined the quality of each article using the tool that assessed the risk of bias and applicability concerns in the key areas of patient selection, index test, reference standard, flow and timing and overall risk of bias. Risk of bias and applicability concerns were classified as low risk, high risk, or unclear. Risk of bias was considered low (high quality evidence) when at least three of four QUADAS-2 criteria were met, and it was considered high (low quality evidence) when two or fewer criteria were met. The results are presented using tables from the QUADAS web site (www.quadas.org). Disagreements were resolved by consensus and when a consensus could not be reached, a third reviewer was consulted. The grading of recommendation, assessment, development, and evaluation (GRADE) was used to synthesize the results for each test by integrating the quality of evidence (high risk/low quality of low risk/high quality) and the magnitude of the effect from statistical pooling to arrive at a recommendation of: strong evidence for use, weak evidence for use, weak evidence against use, strong evidence against use, or inconclusive (Macaskill et al., 2010).

2.4. Data extraction and analysis

The authors extracted data characteristics into a standardized database that included information on study setting, participants, study design, representative spectrum, reference standard, index and comparator test and follow up. Raw data on diagnostic accuracy were extracted and a 2x2 (true positives, false negatives, false positives, true negatives) table was constructed where possible to estimate pooled sensitivity (SN), specificity (SP) and likelihood ratio (LR) for any symptom or sign. When data were incomplete, an attempt was made to contact the study's primary author for clarification. If the primary author did not respond after two attempts, the data from that study were not included in the analysis. The pooled data analysis was performed using Meta-DiSc, version 1.4(Zamora et al., 2006).

The heterogeneity of calculated effect sizes was assessed using chi-square (χ^2) and I-square (I^2) tests. The χ^2 was used to test the hypothesis that two or more population distributions differed from one another at a level greater than chance, with a significant value ($p < 0.05$) indicating heterogeneity. The I^2 value is an estimate of the percentage of total variation that is due to heterogeneity among studies rather than due to chance (Higgins et al., 2003). For the I^2 score, 25% indicates a low level of heterogeneity among effects from primary sources; 50%, moderate; and 75%, high.

2.5. Moderator analysis

Moderator analyses were planned for any pooled result that revealed statistically significant heterogeneity. This would normally occur by stratifying the sample of papers by the moderator variable to determine the degree of difference in calculated effect sizes by level of moderator. If stratifications and re-analysis reduced the heterogeneity within strata to a non-significant level, that variable was deemed a potentially meaningful moderator (Higgins et al., 2003) This is only logically possible where at least 3 primary sources are included in the analysis.

3. Results

Fig. 1 shows the search strategy and results. The initial database search identified 824 articles relevant to CES. After merging of results from all databases, removal of duplicates and screening for titles and abstracts, we retrieved 42 full text articles from the databases and 6 articles from the grey literature. Upon full article review, seven were included in the final pool of which six were retrospective studies (Ahad et al., 2015; Balasubramanian et al., 2010; Domen et al., 2009; Gooding et al., 2013; Raison et al., 2014; Rooney et al., 2009) and one prospective cohort study (Bell et al., 2007). One article (Ahad et al., 2015)

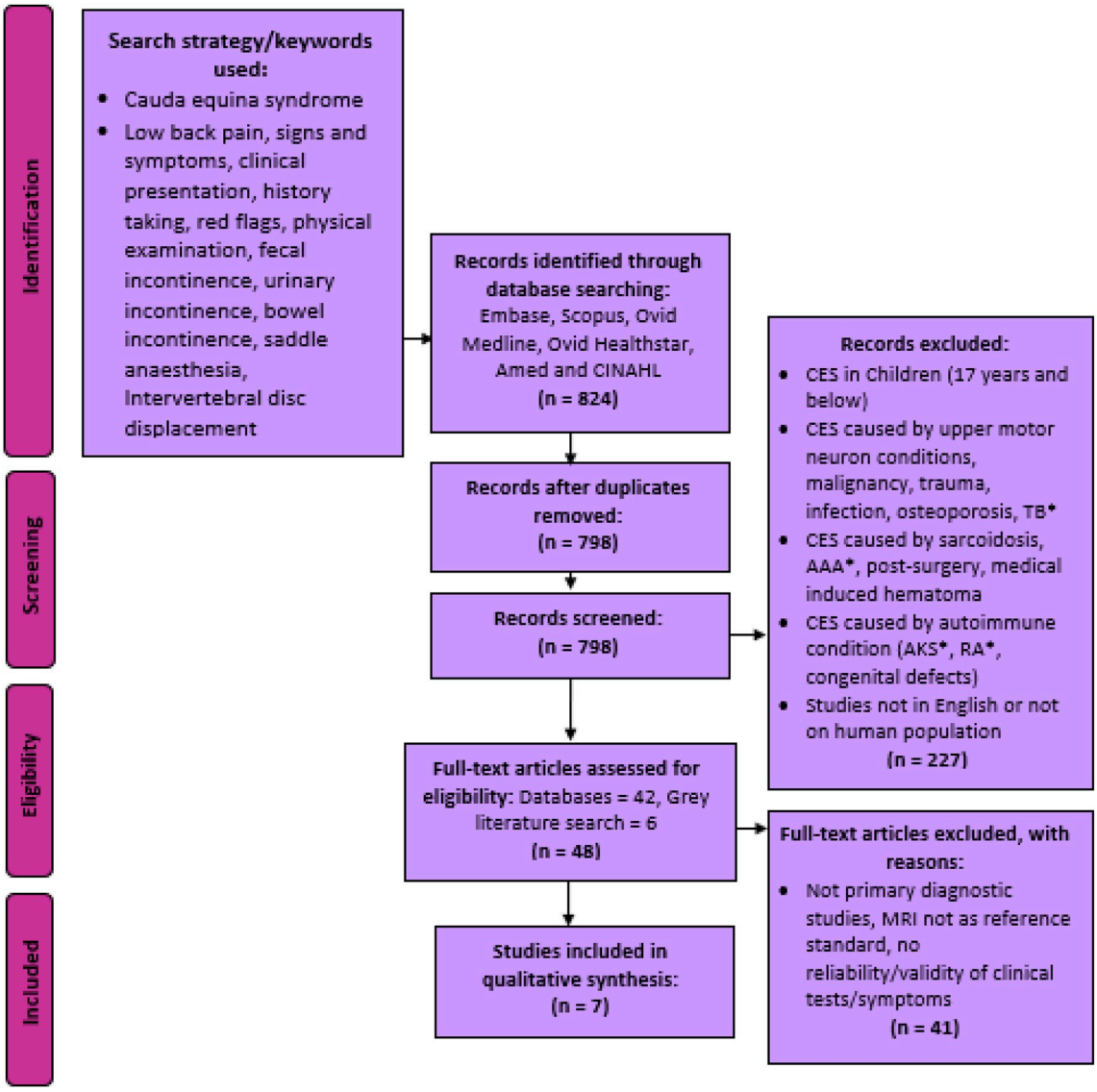


Fig. 1. Prisma flow diagram(Moher et al., 2009).

was not included in the pooling because the published data were inadequate and the primary author did not respond to two contact attempts.

Table 1 shows the individual results of quality assessment for the included studies. All were rated high risk of bias (low quality evidence). The most common potential biases were patient selection and the conduct and interpretation of the index tests. Every included study contained at least one criterion with a rating of 'unclear' indicating deficient reporting to pass judgment on the potential for that bias.

The characteristics extracted for each included study are shown in Table 2. All included studies were conducted in secondary or tertiary care settings: two at the emergency department (Domen et al., 2009; Raison et al., 2014), three at hospital spinal (orthopaedic) units (Ahad et al., 2015; Balasubramanian et al., 2010; Gooding et al., 2013) and two at neurosurgical units (Bell et al., 2007; Rooney et al., 2009). Across all 7 studies, 869 total participants were evaluated. Multiple

Table 1
QUADAS-2.

study	Patient selection	Risk of bias				Applicability concerns		
		Index test	Referencestandard	Flow and timing	Overall risk of bias	Patient selection	Index test	Reference standard
Ahad et al., 2015	☹	☹	?	☹	☹	☺	☺	☺
Balasubramanian et al., 2010	☹	?	?	☹	☹	☺	☺	☺
Bell et al., 2007	☹	☹	☹	?	☹	☺	☺	☺
Domen et al., 2009	☹	?	?	?	☹	☺	☺	☺
Gooding et al., 2013	?	☹	☹	?	☹	☺	☺	☺
Raison et al., 2014	☹	☹	☹	?	☹	☺	☺	☺
Rooney et al., 2009	☹	☹	☹	☹	☹	☺	☺	☺

☺Low Risk; ☹High Risk; ? Unclear Risk.

Table 2
Characteristics of the included studies.

Authors	Setting	Clinical Features	Participants	Study Design	Results	Reference Standard	Symptom/Sign Tests
Ahad et al., 2015	Royal Alexandra Hospital, Paisley, United Kingdom (UK).	Medical notes were reviewed to identify the duration of symptoms, clinical features on examination, result of the scan, the urgency and type of treatment provided	79 consecutive patients 24 males and 55 females	Retrospective study	62 patients (78.5%) showed evidence of lumbar disc pathology on MRI. 45 had disc prolapse, of these 29 occurred at L*4-L5	MRI*	bladder incontinence, saddle anaesthesia, decreased anal tone urinary retention faecal incontinence constipation motor weakness in lower limbs sensory deficit/ numbness/altered sensation.
Balasubramanian et al., 2010	Secondary tertiary referral center for spinal disorders with specialist inin spinal surgical and neurosurgical training	Patients with suspected CES* seen by the on call spinal surgical team.	80 patients age range of patients was between 21 and 90 57% of patients were in the 4th and 5th decades	Retrospective cohort study	18.8% of these patients had CES due to intervertebral disc prolapse confirmed on MRI.	MRI	back pain, unilateral/bilateral leg pain, bladder retention, urinary incontinence, bowel incontinence, saddle sensory deficit, reduced anal tone.
Bell et al., 2007	tertiary referral neurosurgical center. The Department of Clinical Neurosciences, Western General Hospital, Edinburg	All adult patients referred directly from primary care trust with a suspected diagnosis of CES.	23 patients were eligible for inclusion over a 4-month period.	Prospective cohort study	MRI was normal in 10 (43%) of patients. A disc prolapse causing cauda equina distortion was present in 5 (22%).	MRI	urinary retention, urinary frequency urinary incontinence altered urinary sensation altered perineal sensation.
Domen et al., 2009	Secondary tertiary neurology referral center. Hospital Emergency room. Maastricht University Medical Center.	All patients referred because of suspected acute CES.	58 consecutive cases were eligible for inclusion, who had an urgent MRI of the lumbar spine.	Retrospective study	8 patients had cauda compression on MRI 50 patients did not have CES on MRI	MRI	Back pain, unilateral/bilateral leg pain, bladder retention urinary incontinence saddle sensory deficit, reduced anal tone.
Gooding et al., 2013	University teaching hospitawith a spinal unit comprising orthopaedic and neurosurgical expertise.	Clinical presentation appropriate to CES.	57 patients institution Mean age of 45 (range 17–84), male to female ratio of 1:2	Retrospective case note review	13 (23%) patients had confirmed CES on MRI (10 with central disc prolapse, 2 with metastatic disease, 1 with L2 fracture from trauma)	MRI	DRE* (sensation and anal tone) back pain sciatica, sensory change (limbs) weakness (limbs) abnormal reflexes (lower limbs) urinary retention (> 500 ml*) urinary frequency urinary incontinence bowel symptoms (incontinence or constipation)
Raison et al., 2014	Secondary tertiary care setting in an emergency department of a district general hospital.	Adults with low back pain presenting to the emergency department that warranted a referral to the orthopedic team	206 adult patients.	Retrospective cohort study	Of the 206 patients, 32 (15.5%) had confirmed cord/cauda equina compression on MRI	MRI	saddle sensory disturbance bowel/ bladder dysfunction saddle sensory disturbance bowel/ bladder dysfunction
Rooney et al., 2009	Primary or secondary tertiary care from the neurosurgical registrars' daily–update list.	Clinical presentation suggestive of CES.	66 patients with and without abnormal MR imaging admitted to a neurosurgical unit with suspected CES	Retrospective study	34 (52%) of the 66 patients had a relevant abnormality on MRI likely to explain their symptoms. There were no statistical differences in sex, timing of scan and clinical features in the history or examination between patients with and those without structural abnormality. Hospital referred patients were more likely to have a relevant abnormality. Patients without abnormality had a significantly shorter inpatient stay and were marginally younger.	MRI	low back pain sciatica, saddle numbness leg numbness fecal/urinary retention sphincter sensation leg power reflexes saddle numbness to pinprickleg numbness.

Abbreviations*: CES: Cauda Equina Syndrome; MRI: Magnetic Resonance Imaging; DRE: Digital Rectal Examination; ml: milliliters; L: Lumbar segment.

clinical signs and symptoms were evaluated in each study (mean of 8 per study). None of the studies specified the follow-up procedures/pathways for the cohorts.

Seven key clinical signs and/or symptoms were evaluated in enough independent cohorts to allow statistical pooling. These were: saddle anaesthesia, bowel incontinence, reduced anal tone, leg pain, urinary retention, urinary incontinence, and back pain. A description of the various terms used for these key clinical signs and/or symptoms is in Appendix C. Except for reduced anal tone, each of the key clinical signs and symptoms demonstrated heterogeneity for both specificity and sensitivity when the data were pooled, as showed in the results below. Potential sources of heterogeneity could include one or more of the following based on our review: study setting (the patient population is different between secondary and tertiary care settings), spectrum bias (most patients sent to tertiary care have a higher likelihood of having a serious condition than in primary or secondary care), and performance bias (the level of expertise of the attending medical staff differed between the studies included).

- Saddle anaesthesia (4 articles, n = 371)(Balasubramanian et al., 2010; Domen et al., 2009; Raison et al., 2014; Rooney et al., 2009) showed pooled specificity (Sp) of 0.85 (95% CI 0.81 to 0.89); and pooled sensitivity (Sn) of 0.38 (95% CI 0.27 to 0.49) with significant heterogeneity in each (I^2 = 90.3%, χ^2 p < 0.001, and I^2 = 81%, χ^2 p < 0.001, respectively). A moderator analysis of Sn and Sp revealed that one article conducted in a tertiary neurosurgical centre (Rooney et al., 2009) was the outlier, and deeper exploration revealed it was the only one of the four articles for which the outcome assessors were not truly blinded to the condition they were assessing. Removal of the data from this paper left three articles (n = 342) with homogeneity revealing pooled Sp of 0.89 (95% CI 0.85 to 0.92, I^2 = 0%, χ^2 p = 0.58) and pooled Sn of 0.30 (95% CI 0.20 to 0.42, I^2 = 50.3%, χ^2 p = 0.13) (Fig. 2a and b). GRADE Recommendation: weak evidence for use, more specific than sensitive.
- Bowel incontinence (4 articles, n = 217)(13,15–17) pooled Sp was heterogeneous 0.86 (95% CI 0.80 to 0.90, I^2 = 0%, χ^2 p = 0.85) but homogeneous for pooled Sn of 0.19 (95% CI 0.09 to 0.37, I^2 = 67.8%, χ^2 p = 0.03)(Fig. 2c and d). GRADE Recommendation: weak evidence for use.
- Reduced anal tone (3 articles, n = 190)(15–17) pooled results were homogenous: pooled Sp was 0.83 (95% CI 0.76 to 0.88, I^2 = 20.3%, χ^2 p = 0.28) with Sn of 0.30 (95%CI 0.16 to 0.49, I^2 = 18%, χ^2 p = 0.30) (Fig. 2e and f). GRADE Recommendation: weak evidence for use
- Leg pain (5 articles, n = 284)(7,13,15–17) pooled Sp was 0.66 (95% CI 0.59 to 0.73, I^2 = 94.5%, χ^2 p < 0.001) with Sn of 0.43 (95% CI 0.30 to 0.56, I^2 = 67.3%, χ^2 p = 0.02) (Fig. 2g and h). GRADE Recommendation: weak evidence for use
- Urinary retention (5 articles, n = 284)(Balasubramanian et al., 2010; Bell et al., 2007; Domen et al., 2009; Gooding et al., 2013; Rooney et al., 2009) pooled Sp was 0.72 (95% CI 0.65 to 0.79, I^2 = 82%, χ^2p < 0.001) and Sn was 0.25 (95% CI 0.17 to 0.35, I^2 = 0%, χ^2 p = 0.66) (Fig. 2i and j). GRADE Recommendation: weak evidence against use.
- Urinary incontinence (5 articles, n = 284)(Balasubramanian et al., 2010; Bell et al., 2007; Domen et al., 2009; Gooding et al., 2013; Rooney et al., 2009) pooled Sp was 0.70 (95% CI 0.61 to 0.77, I^2 = 76.1%, χ^2 p = 0.002) with Sn of 0.24 (95% CI 0.16 to 0.33, I^2 = 24.2%, χ^2 p = 0.26). (Fig. 2k, l). Study setting, performance and spectrum bias were again all potential explanations for the heterogeneity observed in specificity. The data from one article (Rooney et al., 2009) was a significant outlier for specificity. When this article was removed from the calculation, the pooled Sp of the remaining sources was homogenous, increasing to 0.75 (95% CI 0.67 to 0.83, I^2 = 49.6%, χ^2 p = 0.11). GRADE Recommendation: weak evidence against use.
- Back pain (3 articles, n = 204)(Balasubramanian et al., 2010; Domen et al., 2009; Rooney et al., 2009) pooled Sp was 0.62 (95% CI 0.51 to 0.72, I^2 = 78.8%, χ^2 p = 0.009) with Sn of 0.34 (95% CI 0.26 to 0.42, I^2 = 97.2%, χ^2 p < 0.001) (Fig. 2m, n). GRADE Recommendation: weak evidence for use.

Inadequate data (number of independent studies, published results) were available to pool abnormal reflexes (Gooding et al., 2013), loss of motor function in leg (Gooding et al., 2013), urinary frequency (Domen et al., 2009; Gooding et al., 2013), perineal sensory changes (Gooding et al., 2013), altered urinary sensation (Bell et al., 2007) and sensory changes in the legs (Gooding et al., 2013).

Table 3 presents the raw and pooled data for each of the 7 symptom/sign tests along which includes the true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), SN, SP and LR when compared to MRI. The positive and negative LRs did not reach clinical relevance (> 5. o) for any of the pooled data.

4. Discussion

We described the results of a systematic review and statistical pooling procedure to identify those key clinical signs and symptoms that can be used with confidence to identify red flags associated with CES. To the best of our knowledge, this review is the first to attempt statistical pooling of the diagnostic accuracy of these tests for this purpose.

Fairbank et al. (2011) performed a systematic review thatidentified a number of signs and symptoms of CES and that were evaluated for their diagnostic accuracy in predicting CES against MRI. This current study was able to identity three additional articles which allowed for statistical pooling of data of seven signs and symptoms.

The application and relevance of sensitivity and specificity can assist with clinical decision making using clinical tests that can confirm or refute the presence of a disease or continue the diagnostic process. The definition of sensitivity refers to the ability of the test to correctly identify those patients with the disease and specificity refers to those without the disease (Lalkhen and MCCluskey, 2008). Every sign/symptom evaluated was more specific than it was sensitive. A review of Table 3 shows that pooled Sn of all 7 tests was below 0.50 and the pooled positive likelihood ratio for urinary retention and urinary incontinence was below parity (1.0), suggesting that a positive result on either of these tests leads to a shift in likelihood of CES that is worse than chance. As such, the GRADE recommendation for these two tests, themselves drawn from low quality evidence, are weak against their use in clinical practice. For the other 5 tests, the findings of higher pooled specificity with very poor sensitivity suggests that, when used in isolation, clinicians can expect a larger than desirable incidence of false negatives. However, a positive result on these is unlikely to be a false positive, meaning that those 5 tests could have some clinical utility for identifying CES when it is indeed present. We suspect that considering the results of several of these tests together as a cluster will further improve their clinical utility, though this has yet to be empirically studied. Our findings are in keeping with those of Fairbank et al., (2011), from a qualitative synthesis of existing studies to conclude that no individual symptom or sign from the patient history or clinical examination could be used to diagnose CES. The less condition-specific symptom of back or leg pain were still more specific than sensitive for this condition, overall suggesting that all tests that have been subject to empirical scrutiny so far are at high risk of missing CE compression when it actually exists (false negatives).

Our approach also included moderator analysis offering insight into some potential context-dependent confounders for diagnostic accuracy. All extracted data were collected from a variety of settings. Two of the included studies (Balasubramanian et al., 2010; Rooney et al., 2009) were often found to be outliers in the data analysis, as highlighted

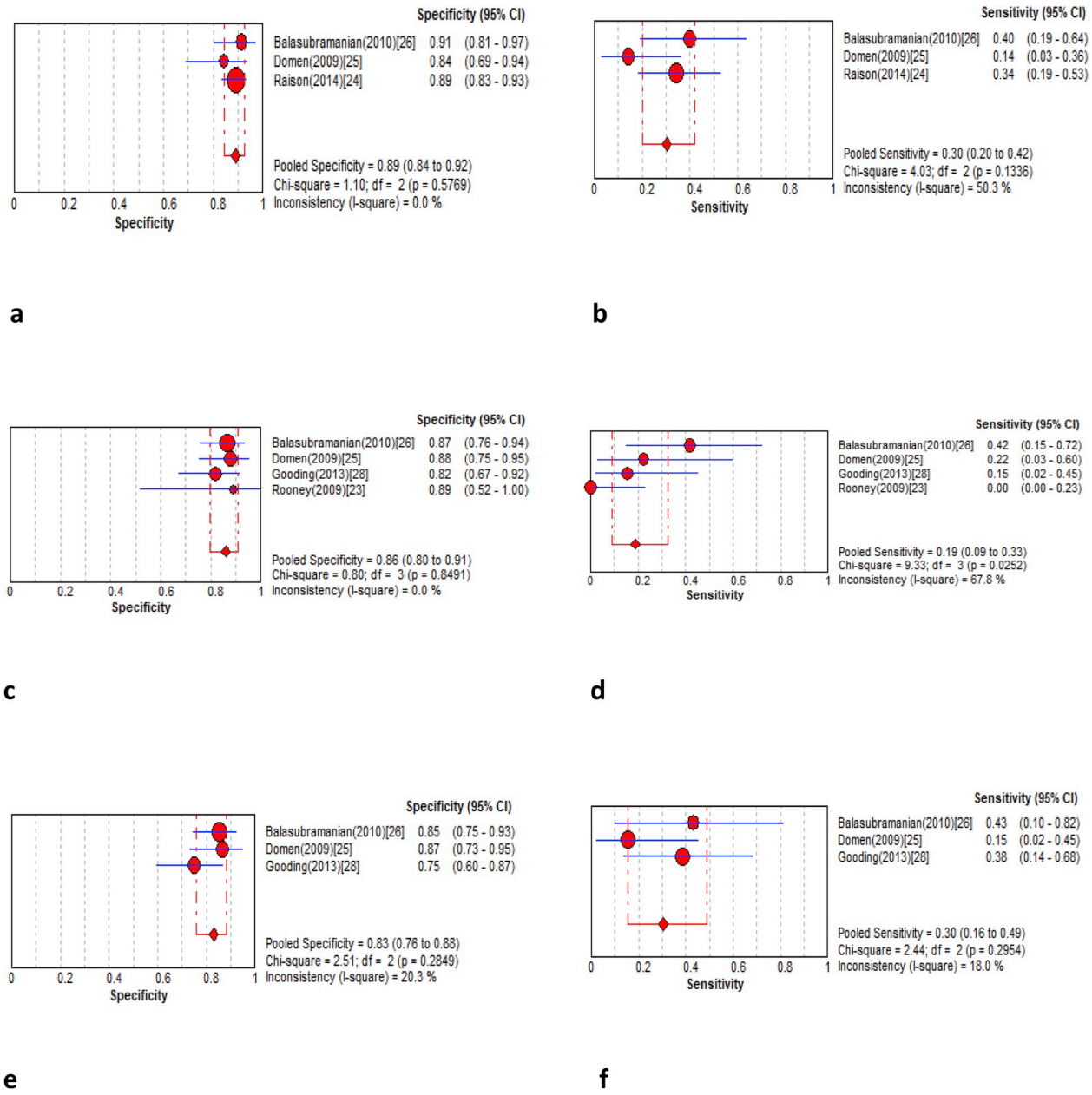


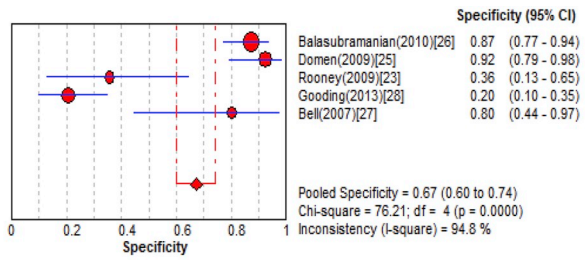
Fig. 2. Forrest plots of specificity and sensitivities for red flags from the included studies, calculated using Meta-Disc version 1.4 (Zamora et al., 2006). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

earlier. A possible explanation for the heterogeneity observed is that they took place in tertiary centres for spinal/orthopaedic surgery (Balasubramanian et al., 2010) and neurosurgery (Rooney et al., 2009) respectively (Fig. 2c and d). A higher prevalence of CES and severity of baseline symptoms is expected in this type of setting as the spectrum of disease that a specialist sees will be significantly different from that of a general practitioner in primary care. It may also be expected that specialists in tertiary care are more adept at promptly identifying certain conditions compared to generalist physicians. It is noteworthy to mention however that in the case of CES, even specialists could not reliably predict its presence or absence in previous studies (Balasubramanian et al., 2010; Bell et al., 2007; Verhagen et al., 2016). These subtle differences in the study setting, population (e.g. baseline disease severity), disease prevalence, grade and experience of the assessors seem to explain the clinical and methodological heterogeneity observed (Guyatt et al., 2011).

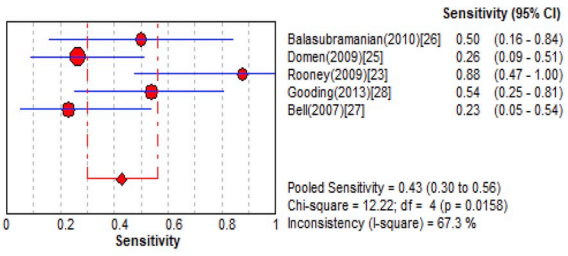
Saddle anaesthesia demonstrated the best pooled diagnostic value of all the identified red flags with a LR+ 2.00. However, this does not reach clinical relevance with a change in probability of about 15% (McGee 2002). Pooled specificity for this test was also at its highest at 0.85. The test with the highest pooled sensitivity was bilateral leg pain at 0.43 and back pain was the symptom with the best LR- at 0.64.

Reduced anal tone was the only pooled sign that demonstrated homogenous results in this review whereas leg pain had the highest statistically significant heterogeneity for sensitivity and specificity (Fig. 2g and h). Moderator analysis for sensitivity of bowel incontinence also revealed a statistically significant heterogeneity. It would appear that the values from the individual studies should not be pooled because point estimate varied widely among the studies, confidence intervals showed minimal overlap, p-values were low and inconsistency values were large (Guyatt et al., 2011; Higgins et al., 2003).

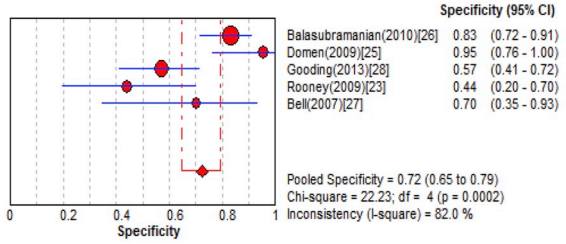
It is important to contrast the results of research studies (in which only a single clinical test is explored at a time) and clinical practice (in which clinicians use multiple sources of information and pattern



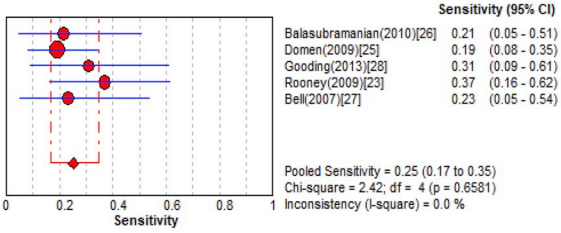
g



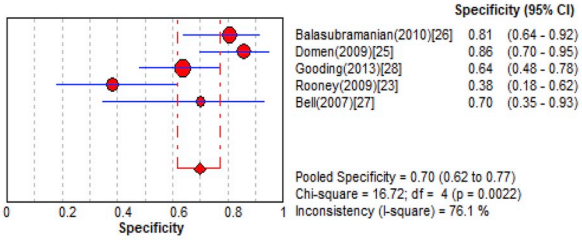
h



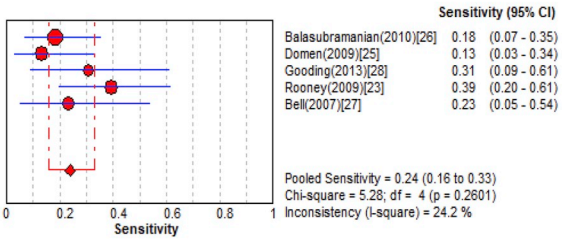
i



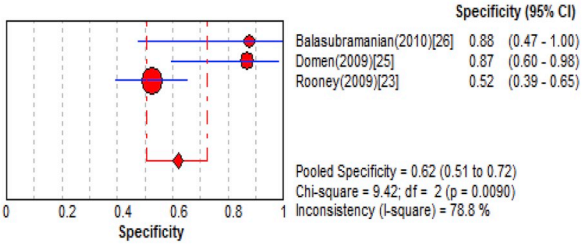
j



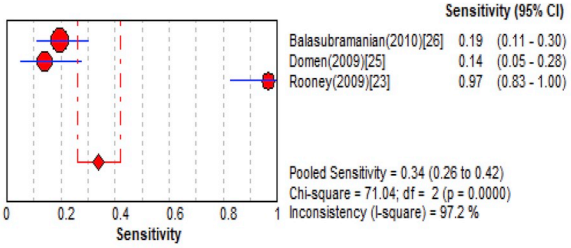
k



l



m



n

Fig. 2. (*continued*)

Table 3
Pooled data for symptom/sign with reported true positives (TP), false positive (FP), false negative (FN), true negative (TN) sensitivity (SN), specificity (SP), positive or negative likelihood ratios (LR+ or -) with 95% confidence intervals (CI), compared to MRI.

Symptom/Sign	Authors	Terminology	TP	FP	FN	TN	SN	SP	LR+	LR-
Saddle Anaesthesia										
	Balasubramanian et al.	Saddle sensory deficit	8	5	12	52	0.4 (0.19–0.64)	0.91 (0.81–0.97)	4.56 (1.69–12.33)	0.66 (0.46–0.95)
	Domen et al.	Saddle anaesthesia	3	6	18	32	0.14 (0.03–0.36)	0.84 (0.69–0.94)	0.90 (0.25–3.25)	1.02 (1.02–0.82)
	Raison et al.	Saddle sensory disturbance	11	19	21	155	0.34 (0.19–0.53)	0.89 (0.84–0.93)	3.15 (1.66–5.97)	0.74 (0.57–0.95)
	Rooney et al.	Numb saddle	10	12	2	5	0.83 (0.52–0.98)	0.29 (0.10–0.56)	1.18 (1.18–0.79)	0.57 (0.13–2.45)
Pooled results via metadisc with 95% CI							**0.38 (0.28-0.49)**	**0.85 (0.81-0.89)**	**2.00 (0.92-4.33)**	**0.80 (0.61-1.05)**
Reduced Anal Tone										
	Balasubramanian et al.	Reduced anal tone, lax anal sphincter	3	10	4	58	0.43 (0.10–0.82)	0.85 (0.75–0.93)	2.91 (1.04–8.16)	0.67 (0.35–1.28)
	Domen et al.	Decreased anal sphincter tonus	2	6	11	39	0.15 (0.02–0.45)	0.87 (0.73–0.95)	1.15 (0.26–5.05)	0.98 (0.75–1.26)
	Gooding et al.	Abnormal tone	5	11	8	33	0.39 (0.14–0.68)	0.75 (0.60–0.87)	1.54 (0.65–3.63)	0.82 (0.52–1.30)
Pooled results via metadisc with 95% CI							**0.30 (0.16-0.49)**	**0.83 (0.76-0.88)**	**1.83 (1.00-3.33)**	**0.90 (0.73-1.12)**
Urinary retention										
	Balasubramanian et al.	Bladder retention	3	11	11	54	0.21 (0.05–0.51)	0.83 (0.72–0.91)	1.27 (0.41–3.95)	0.95 (0.70–1.27)
	Domen et al.	Urinary retention	7	1	30	20	0.19 (0.80–0.35)	0.95 (0.76–0.10)	3.97 (0.52–30.12)	0.85 (0.71–1.02)
	Gooding et al.	Bladder retention	4	19	9	25	0.31 (0.9–0.61)	0.57 (0.41–0.72)	0.71 (0.29–1.72)	1.22 (0.78–1.90)
	Rooney et al.	Retention- urine	7	9	12	7	0.37 (0.16–0.62)	0.44 (0.20–0.70)	0.65 (0.32–1.36)	1.44 (0.75–2.77)
	Bell et al.	Painful and painless urinary retention	3	3	10	7	0.23 (0.50–0.59)	0.70 (0.35–0.93)	0.77 (0.19–3.03)	1.10 (0.66–1.82)
Pooled results via metadisc with 95%CI							**0.25 (0.17-0.35)**	**0.72 (0.65-0.79)**	**0.84 (0.53-1.32)**	**0.99 (0.82-1.20)**
Leg Pain										
	Balasubramanian et al.	Bilateral leg pain	4	11	4	61	0.50 (0.16–0.84)	0.85 (0.74–0.92)	3.27 (1.36–7.90)	0.60 (0.29–1.19)
	Domen et al.	Bilateral symptoms	5	3	14	36	0.26 (0.09–0.51)	0.92 (0.79–0.98)	3.42 (0.91–12.83)	0.80 (0.60–1.06)
	Rooney et al.	Sciatica	7	9	1	5	0.88 (0.47–0.10)	0.36 (0.13–0.65)	1.36 (0.85–2.18)	0.35 (0.05–2.49)
	Gooding et al.	Sciatica	7	35	6	9	0.54 (0.25–0.81)	0.21 (0.10–0.35)	0.68 (0.40–1.14)	2.26 (0.99–5.16)
	Bell et al.	Bilateral sciatica	3	2	10	8	0.23 (0.05–0.54)	0.80 (0.44–0.98)	1.15 (0.24–5.65)	0.96 (0.63–1.48)
Pooled results via metadisc with 95% CI							**0.43 (0.30-0.56)**	**0.66 (0.59-0.73)**	**1.50 (0.80-2.80)**	**0.90 (0.61-1.30)**
Urinary Incontinence										
	Balasubramanian et al.	Bladder incontinence	6	7	27	29	0.18 (0.07–0.35)	0.81 (0.64–0.92)	0.94 (0.35–2.50)	1.02 (0.81–1.27)
	Domen et al.	Urinary incontinence	3	5	20	30	0.13 (0.03–0.34)	0.86 (0.70–0.95)	0.91 (0.24–3.46)	1.01 (0.82–1.25)
	Gooding et al.	Bladder incontinence	4	16	9	28	0.31 (0.09–0.61)	0.64 (0.48–0.78)	0.85 (0.34–2.09)	1.09 (0.71–1.67)
	Rooney et al.	Incontinence-urine	9	13	14	8	0.39 (0.20–0.62)	0.38 (0.18–0.62)	0.63 (0.34–1.16)	1.60 (0.85–3.02)
	Bell et al.	Urinary incontinence	3	3	10	7	0.23 (0.05–0.54)	0.70 (0.35–0.93)	0.77 (0.20–3.03)	1.10 (0.66–1.82)
Pooled results via metadisc with 95% CI							**0.24 (0.16-0.33)**	**0.70 (0.61-0.77)**	**0.76 (0.50-1.13)**	**1.05 (0.92-1.20)**
Bowel Incontinence										
	Balasubramanian et al.	Bowel incontinence	5	9	7	58	0.42 (0.15–0.72)	0.87 (0.76–0.94)	3.10 (1.26–7.66)	0.67 (0.41–1.20)
	Domen et al.	Rectal incontinence	2	6	7	43	0.22 (0.03–0.60)	0.88 (0.75–0.95)	1.82 (0.43–7.61)	0.89 (0.62–1.28)
	Gooding et al.	Faecel incontinence	2	8	11	36	0.15 (0.02–0.45)	0.81 (0.67–0.92)	0.85 (0.20–3.50)	1.03 (0.79–1.36)
	Rooney et al.	Incontinence-Faeces	0	1	14	8	0.000 (0.00–0.23)	0.89 (0.52–0.20)	0.22 (0.01–4.93)	1.14 (0.86–1.50)
Pooled results via metadisc with 95% CI							**0.19 (0.09-0.33)**	**0.86 (0.80-0.91)**	**1.60 (0.66-3.89)**	**0.97 (0.78-1.20)**
Back Pain										
	Balasubramanian et al.	Back pain	14	1	58	7	0.19 (0.11–0.30)	0.88 (0.47–1.00)	1.56 (0.23–10.32)	0.92 (0.69–1.22)
	Domen et al.	Low Back pain	6	2	37	13	0.14 (0.05–0.28)	0.87 (0.60–0.98)	1.05 (0.24–4.64)	0.99 (0.78–1.25)
	Rooney et al.	Low Back pain	29	29	1	32	0.97 (0.83–1.00)	0.52 (0.39–0.65)	2.03 (1.55–2.67)	0.06 (0.01–0.44)
Pooled results via metadisc with 95% CI							**0.34 (0.26-0.42)**	**0.62 (0.51-0.72)**	**1.98 (1.52-2.58)**	**0.64 (0.26-1.60)**

recognition to arrive at a diagnosis). Given the nature of data we extracted, it was not possible to explore the potential clinical utility of a combination or cluster of tests. Intuitively, it would be expected that requiring two or more tests to be positive would further increase its validity. In light of our findings where specificity was already high in some cases, the presence of a single positive sign warrants further diagnostic work-up.

Another contributing factor for the heterogeneity of our results could also be due to the lack of consensus as to how CES is defined in the literature. Since the definition of cauda equina syndrome itself still appears to be ambiguous amongst healthcare practitioners (Fraser et al., 2009), clarifying what CES entails, as well as establishing evidence-based guidelines regarding effective management is essential. Todd and Dickson (2016), have recently identified five characteristic features of CES and separated the disease presentation into four categories in order to aid healthcare providers identify and manage the condition. Greenhalgh et al., (2016) developed a toolkit to facilitate early CES identification, and include the importance of communication and documentation by clinicians, and the empowerment of patients (Greenhalgh et al., 2016). This type of resource helps clinicians use clear, concise and consistent language to ensure that both healthcare practitioners as well as patients understand the gravity of CES, its various signs and symptoms as well as the appropriate action that must be taken (Greenhalgh et al., 2016).

The limitations of this current review were similar to those reported in the review by Fairbank et al., (2011). Small sample sizes and the challenges of conducting robust prospective research may be explained by the rarity of CES (Verhagen et al., 2016), thus making retrospective data collection more convenient.

The following limitations were also noted:

- The data available is generated from secondary and tertiary care settings, making the generalization of the results to primary care settings questionable.
- A lack of a priori study protocol with notable unclear ratings in our quality assessment. Incomplete data records (Ahad et al., 2015; Gooding et al., 2013; Rooney et al., 2009), lack of standardized assessment protocol compared to when a prospective study design is used(Bell et al., 2007).
- Unclear if all included studies consistently adopted the Standards for the Reporting of Diagnostic Accuracy studies (STARD)(Bossuyt et al., 2015) statement in their reporting.

- Possible overestimation of effect from included studies. We were not able to confidently ascertain if reference standard assessors were blinded to the results of the index tests.
- Overall high risk of bias and applicability concerns due to some uncertainties surrounding patient selection, conduct and interpretation of index tests and the reference standard.

Overall, healthcare practitioners may use the results of this study with caution to help guide their clinical reasoning. Given the potential life-changing sequalae of CES, clinicians should have a low threshold when deciding to refer patients for an urgent MRI if cauda equina compromise is suspected.

5. Conclusion

We conclude that the key clinical signs and symptoms commonly used as red flags to screen for CES are not robust enough to diagnose CES on their own as their diagnostic accuracy is poor. That being said, these red flags still remain important clinical signs and symptoms in the suspicion of CES as presently they are the best tools that general healthcare practitioners have to screen for this serious condition.

Presently, MRI remains the investigation of choice to confirm the presence of CES (Ahad et al., 2015; Bell et al., 2007).

Future prospective research enrolling larger cohorts is needed to strengthen evidence in this area (Woodfield et al., 2018). We conclude from our results that existing guidelines need to be updated to reflect:

- Standardized operational clinical signs and symptom definitions and terminology for red flags for CES, which could be used by practitioners in different care settings. It would also allow for improved inter-rater reliability as well as accuracy of pooled results in any future research.
- The need for future research to evaluate red flags as clusters and determine if their diagnostic accuracy is improved when used in this manner rather than used as individual tests(Cook et al., 2018; Underwood, 2009).

Low back pain is the leading cause of global disability(Fraser et al., 2009; Verhagen et al., 2016). As such, individuals with low back pain that may progress to CES need to be made aware of the signs and symptoms of this condition as well as the importance of presenting to the emergency department as soon as these symptoms occur. The informative CES toolkit developed by Greenhalgh et al. (2016) provides a very good resource for such patients. Whilst red flags for CES do not have good overall diagnostic accuracy, recognition and objective assessment by healthcare practitioners remains crucial in identifying which patients should be referred on for an emergency MRI (Greenhalgh et al., 2016).

Declaration of interest

All authors declare that there are no conflicts of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.msksp.2019.05.004.

References

Ahad, A., Elsayed, M., Tohid, H., 2015. The accuracy of clinical symptoms in detecting cauda equina syndrome in patients undergoing acute MRI of the spine. NeuroRadiol. J. 28, 438–442. https://doi.org/10.1177/1971400915598074.

Balasubramanian, K., Kalsi, P., Greenough, C.G., Kuskoor Seetharam, M.P., 2010. Reliability of clinical assessment in diagnosing cauda equina syndrome. Br. J. Neurosurg. 24, 383–386. https://dx.doi.org/10.3109/02688697.2010.505987.

Bell, D.A., Collie, D., Statham, P.F., 2007. Cauda equina syndrome: what is the correlation between clinical assessment and MRI scanning? Br. J. Neurosurg. 21, 201–203. https://dx.doi.org/10.1080/02688690701317144.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L., et al., 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ h5527. https://doi.org/10.1136/bmj.h5527.

Bydon, M., Lin, J.A., De la Garza-Ramos, R., Macki, M., Kosztowski, T., Sciubba, D.M., et al., 2016. Time to surgery and outcomes in cauda equina syndrome: an analysis of 45 cases. World Neurosurg 87, 110–115. https://dx.doi.org/10.1016/j.wneu.2015.12.054.

Cook, C.E., George, S.Z., Reiman, M.P., 2018. Red flag screening for low back pain: nothing to see here, move along: a narrative review. Br. J. Sports Med. 52, 493–496. https://doi.org/10.1136/bjsports-2017-098352.

Domen, P.M., Hofman, P.A., van Santbrink, H., Weber, W.E.J., 2009. Predictive value of clinical characteristics in patients with suspected cauda equina syndrome. Eur. J. Neurol. 16, 416–419. https://doi.org/10.1111/j.1468-1331.2008.02510.x.

Fairbank, J., Hashimoto, R., Dailey, A., Patel, A.A., Dettori, J.R., 2011. Does patient history and physical examination predict MRI proven cauda equina syndrome? Evid. Based Spine Care J. 2, 27–33. https://doi.org/10.1055/s-0031-1274754.

Fraser, S., Roberts, L., Murphy, E., 2009. Cauda equina syndrome: a literature review of its definition and clinical presentation. Arch. Phys. Med. Rehabil. 90https://doi.org/10.1016/j.apmr.2009.03.021. 1964–8.

Germon, T., Ahuja, S., Casey, A.T.H., Todd, N.V., Rai, A., 2015. British Association of Spine Surgeons standards of care for cauda equina syndrome. Spine J. 15, S2–S4. https://doi.org/10.1016/j.spinee.2015.01.006.

Gooding, B.W.T., Higgins, M.A., Calthorpe, D.A.D., 2013. Does rectal examination have any value in the clinical diagnosis of cauda equina syndrome? Br. J. Neurosurg. 27, 156–159. https://doi.org/10.3109/02688697.2012.732715.

Greenhalgh, S., Truman, C., Webster, V., Selfe, J., 2016. Development of a toolkit for early identification of cauda equina syndrome. Prim. Health Care Res. Dev. 17, 559–567. https://doi.org/10.1017/S1463423616000062.

Guyatt, G.H., Oxman, A.D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., et al., 2011. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. J. Clin. Epidemiol. 64, 1294–1302. https://doi.org/10.1016/j.jclinepi.2011.03.017.

Henschke, N., Maher, C.G., Ostelo, R.W.J.G., de Vet, H.C.W., Macaskill, P., Irwig, L., 2013. Red flags to screen for malignancy in patients with low-back pain. Cochrane Database Syst. Rev. CD008686. https://doi.org/10.1002/14651858.CD008686.pub2.

Higgins, J.P.T., Thompson, S.G., Deeks, J.J., Altman, D.G., 2003. Measuring inconsistency in meta-analyses. BMJ 327, 557–560.

Lalkhen, A.G., MCCluskey, A., 2008. Clinical tests:sensitivity and specificity. Critical Care & Pain 8 (6).

Lavy, C., James, A., Wilson-MacDonald, J., Fairbank, J., 2009. Cauda equina syndrome. BMJ 338, 881–884. https://doi.org/10.1136/bmj.b936.

Macaskill, P., Gatsonis, C., Deeks, J.J., Harbord, R.M., Takwoingi, Y., 2010. Chapter 10: Analysing and Presenting Results. In: Deeks, J.J., Bossuyt, P.M., Gatsonis, C. (Eds.), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration Available from. http://srdta.cochrane.org/.

McGee, Steven, 2002. Simplifying likelihood ratios. J. Gen. Intern. Med. 17 (8), 647–650. https://doi.org/10.1046/j.1525-1497.2002.10750.x.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 339https://doi.org/10.1136/bmj.b2535. b2535–b2535.

Raison, N.T.J., Alwan, W., Abbot, A., Farook, M., Khaleel, A., 2014. The reliability of red flags in spinal cord compression. Arch Trauma Res 3, e17850. https://doi.org/10.5812/atr.17850.

Rooney, A., Statham, P.F., Stone, J., 2009. Cauda equina syndrome with normal MR imaging. J. Neurol. 256, 721–725. https://doi.org/10.1007/s00415-009-5003-z.

Todd, N.V., Dickson, R.A., 2016. Standards of care in cauda equina syndrome. Br. J. Neurosurg. 30, 518–522. https://doi.org/10.1080/02688697.2016.1187254.

Underwood, M., 2009. Diagnosing acute nonspecific low back pain: time to lower the red flags? Arthritis Rheum. 60, 2855–2857. https://doi.org/10.1002/art.24858.

Verhagen, A.P., Downie, A., Popal, N., Maher, C., Koes, B.W., 2016. Red flags presented in current low back pain guidelines: a review. Eur. Spine J. 25, 2788–2802. https://doi.org/10.1007/s00586-016-4684-0.

Whiting, P.F., Rutjes, A.W.S., Westwood, M.E., Mallett, S., Deeks, J.J., Reitsma, J.B., et al., 2011. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann. Intern. Med. 155, 529. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.

Woodfield, J., et al., 2018. Understanding cauda equina syndrome: protocol for a UK multicentre prospective observational cohort study. BMJ Open 8, e025230. https://doi.org/10.1136/ bmjopen-2018–025230.

Zamora, J., Abraira, V., Muriel, A., Khan, K., Coomarasamy, A., 2006. Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Med. Res. Methodol. 6, 31. https://doi.org/10.1186/1471-2288-6-31.