ISSN 0008-3194 (p)/ISSN 1715-6181 (e)/2021/186–192/$2.00/©JCCA 2021

Inter-rater reliability of the Quebec Task Force classification system for recent-onset Whiplash Associated Disorders

Yaadwinder Shergill, DC[1,2]
Pierre Côté, DC, PhD[3,4]
Heather Shearer, DC, MSc[3,4]
Jessica J. Wong, DC, MPH[3,4]
Maja Stupar, DC, PhD[5]
Anthony Tibbles, DC[6]
J. David Cassidy, DC, PhD[7]

Purpose: *The inter-rater reliability of the Quebec Task Force (QTF) classification system for Whiplash-Associated Disorders (WAD) remains unknown. Our objective was to determine the inter-rater reliability of the WAD classification between an experienced chiropractic clinician and two chiropractic residents.*

Methods: *We conducted an inter-rater reliability study using baseline clinical data from 80 participants*

Fiabilité inter-évaluateur de la classification établie par le Groupe de travail du Québec sur les troubles associés au coup de fouet cervical d'apparition récente

Objectif : *La fiabilité inter-utilisateur du système de classification des troubles associés au coup de fouet cervical (TACF) établi par le Groupe de travail du Québec (GTQ) demeure inconnue. Notre étude visait à établir la fiabilité inter-évaluateur du système de classification des troubles associés au TACF utilisé par un chiropraticien clinicien d'expérience et deux résidents en chiropratique.*

Méthodologie : *On a effectué notre étude à l'aide de données cliniques de départ sur 80 participants à un*

[1] Department of Health Research Methods, Evidence, and Impact – McMaster University
[2] One Elephant Integrative Health Team, Oakville
[3] Faculty of Health Sciences, Ontario Tech University and Institute for Disability and Rehabilitation Research
[4] Dalla Lana School of Public Health, University of Toronto
[5] Department of Graduate Education and Research, Canadian Memorial Chiropractic College
[6] Clinics, Canadian Memorial Chiropractic College
[7] Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto

Corresponding author: Yaadwinder Shergill, McMaster University Medical Centre, 1280 Main St W. Hamilton, Ontario, Canada L8S 4K1.
Tel: 416-708-2102
E-mail: shergiy@mcmaster.ca

The authors have no disclaimers, competing interests, or sources of support or funding to report in the preparation of this manuscript.

assessed for inclusion in a randomized clinical trial of the conservative management of WAD grades I and II. We reported reliability using Cohen's kappa (k) and 95% confidence intervals (CI).

Results: *The mean duration of WAD symptoms was 7.6 days (s.d.=5.2). In our study, the interrater reliability of the WAD grade classification varied from k=0.04 (95% CI -0.04 to 0.12) to k=0.80 (95% CI 0.67 to 0.94).*

Conclusion: *Inter-rater reliability of the WAD classification varied greatly across raters and may be associated with the experience of the raters and with their understanding of the criteria. Our results suggest that clinicians may benefit from training to standardize how they classify WAD. Furthermore, our results need to be tested in a different sample of patients and with a range of clinicians from different clinical disciplines.*

(JCCA. 2021;65(2):186-192)

KEY WORDS: whiplash-associated disorder, neck pain, classification, Quebec task force, diagnosis

essai clinique, à répartition aléatoire, sur le traitement conservateur du TACF de stades I et II. On a utilisé le coefficient kappa (k) de Cohen et des intervalles de confiance (IC) à 95 % pour évaluer la fiabilité.

Résultats : *La durée moyenne des symptômes du TACF était de 7,6 jours (écart-type :5,2). La fiabilité inter-utilisateur de la classification des TACF a varié de k = 0,04 (IC à 95 % – de 0,04 à 0,12) à k = 0,80 (IC à 95 % de 0,67 à 0,94).*

Conclusion : *La fiabilité inter-utilisateur de la classification des TACF a beaucoup varié d'un évaluateur à l'autre; l'écart pourrait être lié à l'expérience de l'évaluateur et à sa compréhension des critères de classification. Selon les résultats de notre étude, les cliniciens pourraient bénéficier d'une formation servant à normaliser leur méthode de classification des TACF. Nos résultats devraient être confirmés par une autre étude utilisant un autre échantillon de patients et un éventail de cliniciens appartenant à diverses disciplines.*

(JCCA. 2021;65(2): 186-192)

MOTS CLÉS : trouble associé au coup de fouet, douleur cervicale, classification, groupe de travail du Québec, diagnostic

Introduction

Whiplash is an acceleration-deceleration mechanism of energy transfer to the neck. It is the most common injury following motor vehicle collisions, affecting 83% of individuals injured in traffic collisions.[1] Whiplash is associated with clinical symptoms including neck pain, arm pain and paresthesias, dizziness, and psychological distress.[1] These symptoms are collectively known as whiplash associated disorders (WAD).[2] In North America, WAD is common, with an estimated incidence of 600 per 100,000 people.[3-5] The annual economic burden ranges in the billions of dollars depending on the country. In the United States, the burden of whiplash injuries, including medical care, disability, and sick leave is estimated at $3.9 billion (USD) annually.[3] In 2007, the accident benefits paid by Ontario insurers for WAD was reported to be $4.12 billion in Canadian dollars.[6]

The current evidence suggests that 50% of individuals with WAD recover within three to six months of their injury.[7] Of those who report symptoms at one-year post-collision, 30-40% report mild to moderate levels of pain and 10-20% report moderate to severe pain.[7] In 1995, the Quebec Task Force on Whiplash-Associated Disorders proposed a classification system for grading WAD injuries (Table 1).[8] The system classifies patients using clinical data collected during the history and physical examination (including pain, decreased range of motion, point tenderness, neurological signs and fracture or dislocation).[8]

Although the Quebec Task Force classification is commonly used to guide the management of WAD, its inter-rater reliability remains unknown.[9,10] Little is known about the clinical utility of the QTF classification. To our knowledge, there is only one study that supports using the QTF classification system to predict the prognosis of

Table 1.
Quebec Task Force Whiplash Associated Disorders Classification System[8]

Grades	Clinical Presentation
0	No complaint about the neck No physical sign(s)
1	Neck complaint of pain, stiffness, or tenderness only No physical sign(s)
2	Neck complaint AND Musculoskeletal sign(s)*
3	Neck complaint AND Neurological sign(s)^
4	Neck complaint AND Fracture or dislocation

* Musculoskeletal signs include decreased range of motion and point tenderness.
^ Neurologic signs include decreased or absent deep tendon reflexes, weakness, and sensory deficits

WAD within 24 months of the injury' the study suggest that the prognosis worsens with increasing WAD grade.[11] Our objective was to determine the inter-rater reliability of the Quebec Task Force classification system in patients with WAD I and WAD II by comparing the physical examination ratings of an experienced clinician and the chart review ratings of two chiropractic residents.

Methods

Design and study sample

We conducted a chart-based inter-rater reliability study. Our study sample included participants enrolled in a randomized clinical trial (RCT) investigating the effectiveness of conservative management of patients with acute grade I/II WAD.[12] Individuals were eligible for the RCT if they made an insurance claim with a large Canadian automobile insurer between February, 2008 and April, 2011 and resided or worked in the Greater Toronto Area, Mississauga, Burlington, Cambridge, or Kitchener. Participants enrolled in the RCT met the following inclusion criteria: 1) 18 years or older; 2) diagnosed with Grade I or Grade II WAD by the trial coordinator; 3) made an insurance claim for physical injury within 21 days of the traffic collision; 4) reported neck pain greater than or equal to 3 on a 0-10 Numerical Rating Scale; and 5) were able to communicate in English. Excluded individuals were those with: 1) fracture/dislocation of the spine (Grade IV WAD); 2) head trauma; 3) previous whiplash injury within one year; 4) active systemic disease; 5) previous neck surgery; and 6) previous care from a physiotherapist or chiropractor for neck pain in the three months prior to the collision.[10] The study sample for this reliability study included 80 randomly selected charts from potential participants who consented to participate in the RCT.

Data used for the determination of WAD grade

Clinical data used in our reliability study was collected by the same trial coordinator, a chiropractor with nine years of clinical experience, who assessed potential participants for their eligibility to the RCT. The trial coordinator was trained in the grading of WAD. The trial coordinator used standardized forms to collect baseline data and ensured completeness of data collection. The data used to classify WAD grade included: 1) a pain diagram completed by the participant[13]; 2) clinical information describing post-collision symptoms; 3) cervical spine range of motion; 4) results from cervical spine palpation; 5) results from neurological examinations, and 6) neck pain intensity rated by the participants as well as disability measured on the Whiplash Disability Questionnaire[12].

Classification of WAD grade by raters

Our inter-rater reliability study involved three raters: the trial coordinator who initially graded WAD when assessing potential participants for the RCT and two chiropractors with two to three years of experience who were completing their post-graduate residency in Chiropractic Clinical Sciences. Both residents attended a training session delivered by the trial coordinator where the structure of the clinical chart and extraction tables were reviewed. The Quebec Task Force classification system was provided, and five charts were used to pilot the WAD classification. The residents were not provided any training regarding the application of the Quebec Task Force classification because it was assumed to be well understood by the raters who underwent the same education at the same institution.

The WAD grade assigned by the trial coordinator when assessing study participants was used for this study for reliability. The residents received a randomly ordered series of charts and classified WAD grade independently. They

Table 2.
Patient characteristics

Variable	N	Mean	Standard deviation	Minimum	Maximum
Age (years)	80	43.02	13.77	20.00	81.30
Days since injury	80	7.61	5.19	1.00	24.00
Neck pain intensity in past 24 hours (0-10)	80	5.89	1.97	2.00	10.00
Whiplash-related disability score – WDQ (0-130)	79*	51.90	30.00	4.00	116.00

[a] Composite disability score; WDQ: whiplash disability questionnaire
* Data for WDQ score was missing from the sample

were blinded to the WAD grading reported by the trial coordinator and to each other's ratings. Residents provided a WAD grade for the participants once they reviewed the charts.

Sample size

We estimated the sample size according to the method described by Cantor.[14] Based on a desired power of 0.80, alpha level of 0.05 and a null reliability coefficient set at 0.85, a sample size of 69 charts was necessary. However, due to availability of data and to improve precision of our estimates, we used a sample size of 80 participants.

Data analysis

We computed an unweighted kappa (k) statistic and 95% confidence intervals (CI) for each pair of raters (trial coordinator and residents).[15] We performed all statistical analyses using SAS statistical software (Version 9.1; SAS Institute Inc, Cary, NC).

Ethics

The study was approved by the Ethics Institutional Review Board of the Canadian Memorial Chiropractic College and the University Health Network.

Results

Study sample

The mean age of the sample (n=80) was 43.0 years and 75.0% were women. On average, participants were assessed 7.6 days after the collision (Table 2). The mean intensity of neck pain was 5.9/10 (SD = 2.0) and the mean level of disability measured on the Whiplash Disability Questionnaire was 51.90/130 (SD =30.0). The characteristics of the sample from which the participants were selected for our reliability are presented elsewhere.[16]

Table 3.
Distribution of responses A: Rater 1 and Rater 2; B. Rater 1 and Rater 3; C. Rater 2 and Rater 3.

A. Rater 1 and Rater 2		**Rater 1**		
Rater 2		**WAD I**	**WAD II**	**Total**
WAD I	Frequency	24	6	30
	Percent	30.8	7.7	38.5
WAD II	Frequency	1	47	48
	Percent	1.3	60.3	61.5
Total	Frequency	25	53	78
	Percent	32.0	68	100
B. Rater 1 and Rater 3		**Rater 1**		
Rater 3		**WAD I**	**WAD II**	**Total**
WAD I	Frequency	1	0	1
	Percent	1.3	0	1.3
WAD II	Frequency	24	54	78
	Percent	30.3	68.4	98.7
Total	Frequency	25	54	79
	Percent	31.6	68.3	100
C. Rater 2 and Rater 3		**Rater 3**		
Rater 2		**WAD I**	**WAD II**	**Total**
WAD I	Frequency	1	29	30
	Percent	1.3	37.2	38.5
WAD II	Frequency	0	48	48
	Percent	0	61.5	61.5
Total	Frequency	1	77	78
	Percent	1.3	98.7	100

Inter-rater reliability

Rater 1 (trial coordinator of primary RCT) classified 25 participants (31.2%) with WAD I and 55 (68.8%) with WAD II. Rater 2 (chiropractic resident) classified 30 participants (37.5%) with WAD I, 48 (60.0%) with WAD II, and 2 (2.5%) as WAD III. Rater 3 (chiropractic resident) reported classified 1 participant (1.3%) with WAD I, 78 (97.5%) as WAD II, and 1 (1.3%) with WAD III (see Table 3).

The percentage agreement between Rater 1 and Rater 2 was 89.0% and the inter-rater reliability was k=0.80 (95% CI 0.67-0.94). The percentage agreement between Rater 1 and Rater 3 was 69.0% and the inter-rater reliability was k=0.05 (95% CI -0.05-0.16). Finally, the percentage agreement for Raters 2 and 3 was 63.0% and the inter-rater reliability was k= 0.04 (95% CI -0.04 to 0.12). WAD III results were removed as that diagnosis was not present in the original exam diagnoses.

Discussion

We measured the inter-rater reliability of the WAD classification system in a sample of participants with WAD grade I and II. We compared the WAD grades of an experienced chiropractor who examined the potential clinical trial participants with the grades extracted from clinical files by two chiropractic residents with two to three years of clinical experience. Our analysis showed important differences in the inter-rater reliability. In fact, our results suggest that the reliability of a well-known classification system can vary significantly between clinicians with different levels of experience. Stynes[17] performed a systematic review on classification of patients with low back-related leg pain. Of the 22 classification systems investigated, six systems reported reliability with scores ranging from not acceptable to great reliability. Varying levels of experience were cited as a possible reason for the low reliability. It is possible that individuals with greater experience are better at classifying due to more exposure to the classification systems.

The low inter-rater reliability in our study may be attributable to several factors. First, raters likely had a different understanding of the WAD classification system due to varying clinical experiences. Specifically, one rater classified 97.5% of participants as suffering from WAD grade II while the other two raters classified 68.8% and 60% of participants as WAD II. A published systematic review assessing the reliability and validity of clinical prediction rules to screen for neck pain reported that inter-rater reliability may be impacted by raters' backgrounds, experience and training.[18] Second, two raters (chiropractic residents) did not assess the participants in person; they relied on the chart's clinical data to classify WAD, whereas the other rater (trial coordinator) assessed the patients clinically. It is likely that in-person assessments provide information that is not adequately captured by clinical charts and/or WAD classification system such as facial grimaces and may be used when classifying. The WAD classification system is dependent exclusively on clinical signs. Third, the criteria for WAD II [i.e., neck complaint and musculoskeletal sign(s) (decreased range of motion and point tenderness)] may be interpreted with ambiguity. Upon further review one rater required the presence of both "decreased range of motion" and "point tenderness" to classify WAD II, whereas the other rater required only one of the musculoskeletal signs to be present for WAD II. The original classification by Spitzer et al. (1995) does not specify how to operationalize the criteria. A modification to the WAD classification system has been suggested by Hartling *et al*.[11] to distinguish between Grade II cases with normal or limited ranges of motion. Previous research supports poorer recovery in patients with both decreased range of motion and point tenderness.[19]

Strengths and limitations

Our study had strengths. First, WAD was graded independently by each rater. Second, the order of rating by each rater for the 80 participant charts was randomized. Third, we used a large sample size to improve the accuracy of data analyses. However, our study has some limitations. The raters may be limited by their years of clinical experience and the patient populations within their private practices. There was also a small number of raters involved in this study. The second and third raters were restricted to written assessment notes to render their diagnosis. Although all relevant clinical information to make the appropriate diagnosis were reported in the patient records, it is possible that non-verbal behaviors had an indirect impact on clinician ratings. This is a component of the physical examination that may aid in diagnosing the WAD grade had the raters observed the physical interaction.[20] Inter-rater reliability may have been improved had the initial clinician-participant interaction been video

recorded. Lastly, it is possible that the low reliability estimates computed in our study may have been due to the low cell counts used for the kappa calculations.[21] To our knowledge, this is the first study assessing the inter-rater reliability of Quebec Task Force WAD classification. It is possible that the original WAD classification may be used reliably by clinicians. Future research on the inter-rater reliability of the classification system should ensure that clinicians are well-trained in the use of the classification to ensure consistent use of the WAD classification. Future studies could also rely on clinicians with similar levels of experience, however these results may not be generalizable to all clinicians because their expertise and training varies. Our results need to be tested in a different sample of patients and with a range of clinicians from different clinical disciplines. Additionally, future studies are needed to investigate the validity and prognostic value of the WAD classification. Finally, as the WAD classification system is already widely adopted, educational measures are needed to target current students, practitioners, and researchers.

Conclusions

The inter-rater reliability of the WAD classification varied greatly between raters. The inconsistency may be associated with raters' experience and understanding of the WAD criteria. Our results suggest that clinicians may benefit from training with clear operational definitions to improve the reliability of the Quebec Task Force classification of WAD. Further, this study highlights the need for clarity in clinical criteria to ensure consistent use of the classification system.

References

1. Yadla S, Ratliff JK, Harrop JS. Whiplash: diagnosis, treatment, and associated injuries. Curr Rev Musculoskelet Med. 2007; 6(1):65-68. doi: 10.1007/s12178-007-9008-x
2. Sterling M. A proposed new classification system for whiplash associated disorders – implications for assessment and management. Man Ther. 2004;9: 60-70. doi:10.1016/j.math.2004.01.006
3. Holm LW, Carroll LJ, Cassidy D et al. The burden and determinants of neck pain in whiplash-associated disorders after traffic collisions: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine. 2008;33(4S): S52-S59. doi: 10.1097/BRS.0b013e3181643ece
4. Cassidy JD, Carrol LJ, Côté P et al. Effect of eliminating compensation for pain and suffering on the outcome of insurance claims for whiplash injury. NEJM. 2000; 342:1179-1186. doi:10.1056/NEJM200004203421606
5. Versteegen GJ, Kingma J, Meijler WJ et al. Neck sprain after motor vehicle accidents in drivers and passengers. Eur Spine J. 2000; 9:547-552.
6. Financial Services Commission of Ontario. Insurance Bureau of Canada. Submission to the Superintendent, Financial Services Commission of Ontario. Submitted for the Review of Part VI of the Insurance Act. 2008.
7. Carroll LJ, Holm LW, Hogg-Johnson S et al. Course and prognostic factors for neck pain in Whiplash-Associated Disorders (WAD). Results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. Spine. 2008; 33(4S):S83-S92. doi: 10.1016/j.jmpt.2008.11.014.
8. Spitzer WO, Skovron ML, Salmi LR, Cassidy JD et al. Scientific monograph of the Quebec Task Force on whiplash-associated disorders: redefining "whiplash" and its management. Spine. 1995;15(20): 1S-73S.
9. Bussieres AE, Stewart G, Al-Zoubi F et al. The treatment of neck-pain associated disorders and whiplash-associated disorders: a clinical practice guideline. J Manipulative Physiol Ther. 2016; 39(8):523-564. doi: 10.1016/.jmpt.2016.08.007.
10. Financial Services Commission of Ontario. Pre-approved Framework Guideline for Whiplash Associated Disorder Grade I Injuries With or Without Complaint of Back Symptoms. Financial Services Commission of Ontario. 2003.
11. Hartling L, Brison R, Ardern C, Pickett W. Prognostic value of the Quebec classification of whiplash-associated disorders. Spine. 2001;26: 36-41. 10.1097/00007632-200101010-00008.
12. Côté PC, Cassidy JD, Carette S et al. Protocol of a randomized controlled trial of the effectiveness of physician education and activation versus two rehabilitation programs for the treatment of whiplash-associated disorders: the University Health Network Whiplash Intervention Trial. Trials. 2008; 9(75). doi: 10.1186/1745-6215-9-75.
13. Southerst D, Stupar M, Cote P et al. The reliability of measuring pain distribution and location using body pain diagrams in patients with acute whiplash associated disorders. J Manipulative Physiol Ther. 2013;36(7): 295-402. doi: 10.1016/j.jmpt.2013.05.023.
14. Cantor A. B. Sample-size calculation for Cohen's kappa. Psychol Method. 1996;1: 150-153.
15. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86: 420-428.
16. Côté P, Boyle E, Shearer HM, Stupar M, Jacobs C, Cassidy JD, Carette S, van der Velde G, Wong JJ, Hogg-Johnson S, Ammendolia C, Hayden JA, van Tulder M, Frank JW. Is a government-regulated rehabilitation guideline more

effective than general practitioner education or preferred-provider rehabilitation in promoting recovery from acute whiplash-associated disorders? A pragmatic randomised controlled trial. BMJ Open. 2019;9(1): e021283. doi: 10.1136/bmjopen-2017-021283.
17. Stynes S, Konstantinou K, Dunn KM. Classification of patients with low backrelated leg pain: a systematic review. BMC Musculoskel Dis. 2016;17:226. doi: 10.1186/s12891-016-1074-z.
18. Moser N, Lemeunier N, Southerst D, Shearer H, Murnaghan K, Sutton D, Cote P. Validity and reliability of clinical prediction rules used to screen for cervical spine injury in alert low-risk patients with blunt trauma to the neck: part 2. A systematic review from the Cervical Assessment and Diagnosis Research Evaluation (CADRE) Collaboration. Eur Spine J. 2018;27(6):1219-1233. doi: 10.1007/s00586-017-5301-6.
19. Sterling M, Carroll LJ, Kasch H, Kamper SJ, Stemper B. Prognosis after whiplash injury: where to from here? Discussion paper 4. Spine. 2011;36(25 Suppl):S330-S334. doi: 10.1097/BRS.0b013e3182388523.
20. Côté P, Shearer H, Ameis A, and the OPTIMa Collaboration. Enabling recovery from common traffic injuries: a focus on the injured person. UOIT-CMCC Centre for the Study of Disability Prevention and Rehabilitation. January 31, 2015. https://www.fsco.gov.on.ca/en/auto/documents/2015-cti.pdf
21. Cicchetti DV, Feinstein AR: High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43: 551-558. doi: 10.1016/0895-4356(90)90159-M.