J Shoulder Elbow Surg (2016) 25, 1577–1582

ELSEVIER

Journal of Shoulder and Elbow Surgery

www.elsevier.com/locate/ymse

High inter-rater reliability, agreement, and convergent validity of Constant score in patients with clavicle fractures

Ilija Ban, MD[a,*], Anders Troelsen, DMSc[a], Morten Tange Kristensen, PhD[b]

[a]*Department of Orthopedic Surgery, Copenhagen University Hospital Hvidovre, Copenhagen, Denmark*
[b]*Physical Medicine and Rehabilitation Research–Copenhagen, Department of Physical Therapy, Copenhagen University Hospital Hvidovre, Copenhagen, Denmark*

Background: The Constant score (CS) has been the primary endpoint in most studies on clavicle fractures. However, the CS was not developed to assess patients with clavicle fractures. Our aim was to examine inter-rater reliability and agreement of the CS in patients with clavicle fractures. The secondary aim was to estimate the correlation between the CS and the Disabilities of the Arm, Shoulder and Hand score and the internal consistency of the 2 scores.

Methods: On the basis of sample sizing, 36 patients (31 male and 5 female patients; mean age, 41.3 years) with clavicle fractures underwent standardized CS assessment at a mean of 6.8 weeks (SD, 1.0 weeks) after injury. Reliability and agreement of the CS were determined by 2 raters. The interclass correlation coefficient ($ICC_{2,1}$), standard error of measurement, minimal detectable change, Cronbach α coefficient, and Pearson correlation coefficient were estimated.

Results: Inter-rater reliability of the total CS was excellent (interclass correlation coefficient, 0.94; 95% confidence interval, 0.88-0.97), with no systematic difference between the 2 raters ($P = .75$). The standard error of measurement (measurement error at the group level) was 4.9, whereas the minimal detectable change (smallest change needed to indicate a real change for an individual) was 13.6 CS points. The internal consistency of the 10 CS items was good, with a Cronbach α of .85, and we found a strong correlation ($r = -0.92$) between the CS and Disabilities of the Arm, Shoulder and Hand score.

Conclusions: The CS was found to be reliable for assessing patients with clavicle fractures, especially at the group level. With high inter-rater reliability and agreement, in addition to good internal consistency, the standardized CS used in this study can be used for comparison of results from different settings.

Level of evidence: Basic Science Study; Validation of Outcomes Instruments or Classification Systems

Keywords: Inter-rater reliability; agreement; Constant score; internal consistency; convergent validity; clavicle fractures

Despite clavicle fracture being a frequent injury, its optimal treatment is still debatable. Functional outcome, primarily assessed with the Constant score (CS), has been the single most important endpoint in recent studies performed on clavicle fractures.[1] The CS was originally designed to assess shoulder disorders in general by combining subjective and objective

The Danish Data Protection Agency approved data collection. Registration and written patient consent were not necessary according to the Ethics Committee of the Capital Region, Denmark (H-6-2014-FSP-003).

*Reprint requests: Ilija Ban, MD, Department of Orthopaedic Surgery, Copenhagen University Hospital Hvidovre, Kettegaard Allé 30, DK-2650 Hvidovre, Denmark.

E-mail address: Ilija.ban@gmail.com (I. Ban).

1058-2746/$ - see front matter
http://dx.doi.org/10.1016/j.jse.2016.02.022

measurements and was not specifically designed to address the outcome of clavicle fracture patients. Several studies testing the psychometric properties of the original CS found good correlation between the CS and several other scales as well as acceptable responsiveness for some specific shoulder disorders other than clavicle fractures.[20] Reliability studies overall found good inter-rater and intrarater reliability of the CS, but these findings have been questioned because the studies probably were biased by lack of standardization.[5] Lack of standardization of the CS has been a major concern and led to a review and guidelines published in 2008.[7] Despite addressing several of the flaws concerning the original work, the new recommendations did not provide a standardized protocol to assess the CS. Therefore, a standardized CS protocol was proposed in a recent work by Ban et al.[2] The inter-rater and intrarater reliability and responsiveness of this protocol were recently established for patients with subacromial impingement syndrome, as well as for the evaluation of the reliability and agreement of 2 strength devices.[6,14,17] However, the psychometric properties of the CS have never been tested in patients with clavicle fractures.

Another often used score to assess patients with clavicle fractures is the Disabilities of the Arm, Shoulder and Hand (DASH) score.[1] The DASH score is a self-administered questionnaire that the patient fills out independently from a rater. Because the CS is more time-consuming for health professionals than the DASH score when used in daily clinical practice, it would be more practical if patients with clavicle fractures could be assessed using the DASH score.

The primary aim of this study was to test the inter-rater reliability and agreement of the newly standardized CS protocol in patients with clavicle fractures. The secondary aims were to test the internal consistency of the CS and DASH score and the convergent validity between the two. We hypothesized that high inter-rater reliability would be achieved using the newly standardized protocol for the CS.[2]

Materials and methods

This study was conducted as a prospective cross-sectional study comparing results obtained using the CS measured by 2 independent raters on the same day. Adult patients (aged >18 years) with a clavicle fracture were included consecutively within 2 weeks after injury from May 2014 to June 2015. To have a representative group of patients, all types of clavicle fractures regardless of anatomic localization and fracture complexity were included. Patients who underwent surgery and those with a displaced midshaft clavicle fracture who were enrolled in an ongoing study (n = 15) were excluded from this study. Furthermore, patients who were not able to read and understand Danish (n = 7) were excluded. At the time of enrollment, the patients were informed about the study before informed consent was obtained.

The CS is a multi-item 0- to 100-point score (high scores indicate a high level of function) with 10 items, half of which are subjectively measured (parts A and B [0 to 35 points]) and the other half objectively (parts C and D [0 to 65 points]).[7] Four of the items are rated on a continuous visual analog scale, with two of these converted to an ordinal scale. The objective strength part is measured on a continuous scale with a maximum of 25 CS points, whereas the remaining items are rated on an ordinal scale. We used a standardized Danish protocol (also published in English) for the CS.[2] Patients were tested 6 to 8 weeks after the time of the primary injury to the clavicle. At this time point after trauma, we expected that the patients were not in a steady state of remission. Thus, we expected a larger variation in CS outcomes compared with, for example, 6 months after injury, where we expected a ceiling effect. Two independent raters (rater A and rater B) assessed each patient on the same day with a minimum of a half hour between the assessments. The 2 raters were blinded to each other's ratings until the end of the study. Randomization as to which rater would start testing was performed at the time of enrollment by computer randomization to ensure that each rater would be the first rater for half of the assessments. The raters (one physiotherapist [M.T.K.] and one orthopedic surgeon [I.B.]) were both experienced in using the CS. We used a long goniometer to assess range of motion and the IsoForce Control dynamometer (Medical Device Solutions, Oberburg, Switzerland) to measure the maximal shoulder strength.[14] The time to administer the test is between 10 and 15 minutes, and it takes approximately 2 minutes to calculate the score.[25]

The DASH questionnaire, designed to evaluate a nonspecific disorder of the upper extremity, consists of 30 items divided into 3 overall categories: physical function, symptoms, and social function. Each item is answered using a 5-point ordinal scale.[11] Similar to the CS, the overall DASH score ranges from 0 to 100, but the interpretation is inverted because 0 represents normal function and 100 represents a dysfunctional upper extremity. The DASH questionnaire was filled out by the patient in the period between the 2 CS assessments. The time to administer the test is between 5 and 7 minutes, and it takes approximately 3 minutes to calculate the score.[25] Furthermore, the preinjury DASH score was assessed by the patient at the time of enrollment.

Sample size was based on estimates provided by Walter et al.[24] With an expected inter-rater correlation coefficient of 0.9 and an acceptable reliability of 0.7, with $\alpha = .05$ and β (type II error) = 0.2, a sample of 18 patients was needed. We chose to include a sample size of double this number to increase the precision of the estimate.

Descriptive statistics were used for demographic variables, with data reported as either numbers with percentages, means with standard deviations, or medians with 25% to 75% quartiles. The depth of the scores (both CS and DASH score) and of single items of the CS was estimated by calculating the percentage of the lowest and highest score. A floor or ceiling effect was present if 15% of patients or more scored the lowest or highest possible score.[16] Inter-rater reliability was calculated using the interclass correlation coefficient (ICC) type 2,1 ($ICC_{2,1}$; ie, 2 way, single measures, absolute agreement), with a 95% confidence interval (CI).[3] Systematic bias of obtained CSs by the 2 raters was examined with the paired *t* test. Agreement at the group level was calculated using the standard error of measurement (SEM), defined as $SEM = SD \times \sqrt{(1 - ICC)}$, with SD representing the standard deviation of the total CS assessments from both raters.[8] The SEM was also used to determine the minimal detectable change (MDC), defined as $MDC = 1.96 \times \sqrt{2} \times SEM$, indicating the smallest change needed to indicate a real change for an individual patient.[4] In addition, we calculated SEM as a percentage ($SEM\% = [SEM/Mean_{Total\_CS}] \times 100$; where $Mean_{Total\_CS}$ indicates the mean of all CS measurements by both raters) and MDC as a percentage ($MDC\% = [MDC/Mean_{Total\_CS}] \times 100$). A Bland-Altman plot with 95% CI levels was used to illustrate the magnitude of agreement between inter-rater values. The inter-rater reliability was estimated