REVIEW ARTICLE

Shoulder & Elbow

Shoulder & Elbow
2018, Vol. 10(4) 274–284

DOI: 10.1177/1758573218774326
journals.sagepub.com/home/sel
SAGE

The reliability and validity of goniometric elbow measurements in adults: A systematic review of the literature

Suzanne F van Rijn[1], Elisa L Zwerus[2], Koen LM Koenraadt[1], Wilco CH Jacobs[1], Michel PJ van den Bekerom[3] and Denise Eygendaal[1,2]

Abstract

Background: The universal goniometer is a simple measuring tool. With this review we aimed to investigate the reliability and validity of the universal goniometer in measurements of the adults' elbow.

Methods: Preferred Reporting Items for Systematic reviews and Meta-Analysis guidelines were followed and our study protocol was published online at PROSPERO. A literature search was conducted on relevant studies. Methodological quality was assessed using the Quality Appraisal of Diagnostic Reliability (QAREL) scoring system.

Results: Out of 697 studies yielded from our literature search, 12 were included. Six studies were rated as high quality. The intrarater reliability intraclass correlation coefficient ranged from 0.45 to 0.99, the interrater reliability ranged from intraclass correlation coefficient 0.53–0.97. One study providing instructions on goniometric alignment did not find a difference in expert versus non-expert examiners. Another study in which examiners were not instructed found a higher interrater reliability in expert examiners. One study investigating the validity of the goniometer in elbow measurements found a maximum standard error of the mean of 11.5° for total range of motion.

Discussion: Overall, the studies showed high intra- and interrater reliability of the universal goniometer. The reliability of the universal goniometer in non-expert examiners can be increased by clear instructions on goniometric alignment.

Keywords

elbow, universal goniometer, goniometry, reliability, validity, 'range of motion'

Date received: 31st December 2017; accepted: 8th April 2018

Background

A patient's ability to perform daily activities such as eating, combing hair, writing and using a PC is highly dependent on the range of motion (ROM) of the elbow. A restricted elbow ROM can result in a serious disability. According to literature on the elbow ROM, in a healthy person values for flexion lie between 130° and 154° and extension between −6° and 11°. Pronation varied from 75° to 85° and supination from 80° to 104°.[1–7] A decrease in ROM can be an indicator of chronic and progressive pathology, such as heterotopic ossifications, osteoarthritis, loose bodies, chondromalacia, valgus extension overload syndrome and osteochondritis dissecans.[8,9] But also in traumatic injuries assessment of elbow ROM can be important, for example an unrestricted elbow extension can rule out a fracture without an X-ray.[10,11] The degree of limitation in elbow ROM can also be used as an indicator of the impact of the disease in daily activities. Daily activities can be performed with an elbow extension

[1]Orthopedic Surgery Department, Amphia Hospital, Breda, the Netherlands
[2]Department of Orthopedic Surgery, Academic Medical Centre, Amsterdam, the Netherlands
[3]Department of Orthopedic Surgery, Shoulder and Elbow Unit, Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands

Corresponding author:
Elisa L Zwerus, Department of Orthopedic Surgery, Academic Medical Centre, Meibergdreef 9, Amsterdam 1105 AZ, the Netherlands.
Email: elisazwerus@gmail.com

restriction of 30° and minimal flexion of 130°, in combination with 50° of pronation and supination;[12] however, some activities as handling a cell phone require more mobility.[13,14] Probably the most important reason to measure the elbow ROM is to closely monitor disease progression or treatment effectiveness.

The universal goniometer (UG) is a simple measuring tool, which is frequently used by many different health care professionals such as physiotherapists, general practitioners and orthopaedic surgeons. Other less studied modalities to measure the elbow ROM include the use of photography, movies, a smartphone application and visual estimation.[15–20] To appreciate the 'true value' of measurements, using the UG, it is important to know its validity and intra- and interrater reliability.

General assumption is that the reliability of the UG is higher when used by an experienced tester and interrater variations are smallest when a standardized measuring method is used.[21,22] Many physicians and physiotherapists use the UG without using an identical measuring protocol if any available.

This systematic literature review investigated the reliability and validity of the UG in the measurement of elbow ROM in adults.

Methods

This systematic review was conducted according to the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement.[23] The study protocol was published online at the PROSPERO International prospective register of systematic reviews (http://www.crd.york.ac.uk/PROSPERO) under registration number CRD42016043760.

Search strategy

A comprehensive electronic literature search was conducted in collaboration with an experienced clinical librarian on relevant articles published from the earliest year until October 2017 in the following databases: EMBASE, MEDLINE ovid, Web of science, Scopus, Cochrane, PubMed publisher, Cinahl EBSCO and Google scholar. The search terms (and synonyms) were 'elbow', 'goniomet', 'range of motion', 'accuracy', 'reliability', 'validity' and 'inter/intra observer'. The reference lists of included articles were manually checked for potentially relevant articles.

Only studies investigating the reliability and/or validity of the UG in elbow measurements in human adults were included. ROM included flexion, extension, pronation, supination and/or carrying angle. Exclusion criteria were a language other than English or Dutch, subjects under the age of 18, animals, full text not available, a different measuring tool than the UG.

Study selection

Two authors (SFR and WCHJ) independently screened all titles and abstracts yielded by the search to identify relevant studies meeting the inclusion and exclusion criteria. The authors were not blinded for author and affiliation names of these studies. Then both authors assessed the full text of the selected articles. Afterwards, the reviewers compared their results, in case of differences they discussed until agreement was reached.

Quality assessment

The quality of the included articles was assessed by two authors (SFR and KLMK) using the QAREL scoring system.[24] This tool scores the articles in their sampling bias: (1) the representativeness of subjects and raters, (2) rater blinding, (3) order of examination, (4) suitability of the time interval, (5) applied and interpreted appropriately and (6) statistical analysis. The maximum score is 11. A study was considered having a high quality when it scored >60% and low quality when scored <60%. This cut-off point has been used in several previous studies.[25–27]

Data extraction

The following data were extracted from all of the included articles by three authors independently (SFR, ELZ and KLMK): (1) population (healthy subjects or symptomatic patients); (2) number of subjects included; (3) movement measured (flexion, extension, pronation, supination or carrying angle); (4) active or passive ROM; (5) if bony landmarks were used or defined prior to the measuring; (6) validity; (7) intra- and interobserver reliability and (8) information about the examiner (profession and/or level of experience in goniometry).

Data analysis

Data analysis was performed by two independent authors (SFR and ELZ) using Microsoft Excel 2010 (Microsoft Corp. Washington, USA). Intraclass correlation coefficient (ICC) less than 0.40 was considered poor, between 0.40 and 0.59 fair, between 0.60 and 0.74 good and >0.75 excellent.[28]

Results

Study selection

A total of 1386 articles were found. After removal of duplicates 697 articles remained. The titles of the 697 articles were screened and 60 articles were selected as

potentially relevant. After reviewing abstracts and/or full text, 48 articles were excluded for various reasons, such as review articles, subjects were children, full text not available, a language other than English or Dutch or the use of a measuring device other than the UG. Twelve articles were finally included for data extraction.[3,6,7,21,29–36] Figure 1 shows the PRISMA flow chart.

Quality assessment

The QAREL checklist showed a high quality (score > 60%) in six out of 12 studies;[7,21,29–31,36] all other studies were of low quality. Most of the studies rated as low quality did not blind (or did not mention to blind) the raters to the findings of other raters[3,6,32–34] or their own prior findings.[3,6,21,30,33] An overview of all the QAREL scores is presented in Table 1.

Included studies

Three studies tested the UG on symptomatic patients,[21,31,36] seven studies used healthy volunteers,[3,6,7,32–35] two studies included both healthy subjects and symptomatic patients.[29,30] Together a number of 376 participants were included. A study by Low[34] only included one subject; however, this study used 50 raters. Rothstein et al.[36] used 12 raters with 12 subjects, all other studies used five or less raters with 23 up to 50 subjects. The number of measurements in all studies was two or three, the interval varied from consequently to four weeks apart. The age varied from 18 to 85 years. Nine studies tested elbow flexion, eight studies extension, five studies pronation and supination. Chapleau et al.[6] also added the carrying angle. Nine studies performed the measurements during active ROM, two during passive ROM[31,36] and one study measured both active and passive.[7] In one study the arms of the subjects were in a fixed position.[34] In four studies the bony landmarks used for the measurements were defined.[6,7,33,35] Two studies investigated the difference between expert and non-expert examiners. Armstrong et al.[21] found no differences between expert and non-expert examiners; Blonna et al.[29] found a slightly lower reliability in non-expert examiners. Cimatti et al.[30] compared injured to non-injured subjects, showing similar interrater reliability for pronation and supination. Characteristics of included studies are presented in Table 2.

Statistical analysis of results

Our intention was to perform heterogeneity analysis and, if applicable, meta-analysis on the included studies. Most studies use the ICC to express the interrater and intrarater reliability. There are several different methods to compute the ICC, for example measuring ICC on single or average values.[37,38] We attempted to determine the method of ICC calculation in every study; however, this was not clear in all studies. Also, some studies presented an ICC range instead of a fixed number. Pooling results is inappropriate in this case.[39] Besides statistical

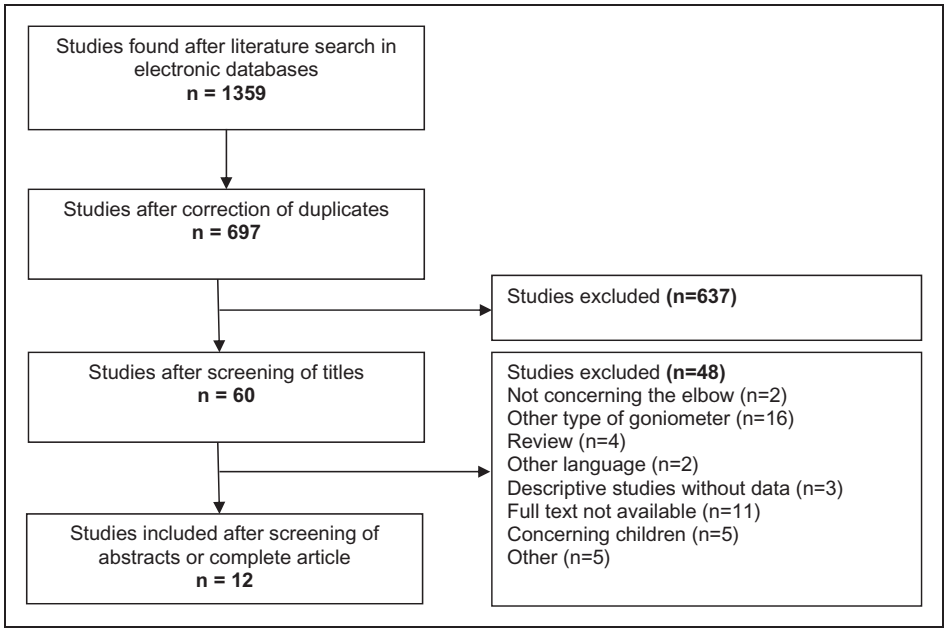

Figure 1. PRISMA flow chart of study selection.

Table 1. QAREL scores.

		1 Armstrong et al.[21]	2 Blonna et al.[15]	3 Chapleau et al.[6]	4 Cimatti et al.[30]	5 Fieseler et al.[3]	6 Flowers et al.[31]	7 Gajdosik[32]	8 Goodwin et al.[33]	9 Low[34]	10 Petherick et al.[35]	11 Rothstein et al.[36]	12 Zwerus[7]
1	Was the test evaluated in a sample of subjects who were representative of those whom the authors intended the results to be applied?	1	1	1	1	1	1	1	1	1	1	1	1
2	Was the test performed by raters who were representative of those to whom the authors intended the results to be applied?	1	1	1	0	1	1	Uncl	1	1	Uncl	1	1
3	Were raters blinded to the findings of other raters during the study?	1	1	0	1	N/A	1	N/A	Uncl	Uncl	1	1	1
4	Were raters blinded to their own prior findings of the test under evaluation?	0	1	0	0	Uncl	1	1	Uncl	1	1	1	1
5	Were raters blinded to the results of the accepted reference standard or disease status for the targeted disorder (or variable) being evaluated?	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

(continued)

Table 1. Continued.

		1 Armstrong et al.[21]	2 Blonna et al.[15]	3 Chapleau et al.[6]	4 Cimatti et al.[30]	5 Fieseler et al.[3]	6 Flowers et al.[31]	7 Gajdosik[32]	8 Goodwin et al.[33]	9 Low[34]	10 Petherick et al.[35]	11 Rothstein et al.[36]	12 Zwerus[7]
6	Were raters blinded to clinical information that was not intended to be provided as part of the testing procedure or study design?	1	Uncl	N/A	Uncl	Uncl	Uncl	N/A	N/A	N/A	N/A	Uncl	N/A
7	Were raters blinded to additional cues that were not part of the test?	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
8	Was the order of examination varied?	1	0	0	1	0	1	Uncl	Uncl	0	Uncl	Uncl	1
9	Was the stability of the variable being measured taken into account when determining the suitability of the time interval between repeated measures?	1	1	N/A	1	1	1	1	1	N/A	N/A	1	1
10	Was the test applied correctly and interpreted appropriately?	1	1	1	1	1	1	1	0	1	1	1	1
11	Were appropriate statistical measures of agreement used?	1	1	1	1	1	Uncl	1	0	0	1	1	1
	Total	**8**	**7**	**4**	**6**	**5**	**7**	**5**	**3**	**4**	**5**	**7**	**8**
	Score (%)	73	64	36	55	45	64	45	27	36	45	64	73

QAREL:Quality Appraisal of Diagnostic Reliability.

Table 2. Study characteristics.

	Population			Measurements						Intrarater		Interrater	
Study	N	Healthy/ symptomatic	Age (years)	Active/ Passive	Fl	Ex	Pro	Sup	CA	Trials (n)	Interval	Raters (n)	Expert/ non-expert
Armstrong et al.[21]	38	Symptomatic	N/A	Active	+	+	+	+	–	2	Same day	5	Both
Blonna et al.[15]	50	Both	18–85	Active	+	+	–	–	–	N/A	N/A	4	Both
Chapleau et al.[6]	51	Healthy	19–50	Active	+	+	–	–	+	3	Consecutively	2	N/A
Cimatti et al.[30]	33	Both	18–70	Active	–	–	+	+	–	3	Consecutively	2	Expert
Fieseler et al.[3]	47	Healthy	18–25	Active	+	+	–	–	–	3	Week	N/A	N/A
Flowers et al.[31]	30	Symptomatic	21–79	Passive	–	–	+	+	–	2	Same day	3	Expert
Gajdosik[32]	31	Healthy	19–40	Active	–	–	+	+	–	3	Consecutively	1	N/A
Goodwin et al.[33]	23	Healthy	18–31	Active	+	+	–	–	–	2	Month	3	Expert
Low[34]	1	Healthy	N/A	Active	+	–	–	–	–	N/A	N/A	50	Expert
Petherick et al.[35]	30	Healthy	20–28	Active	+	+	–	–	–	3	Consecutively	2	N/A
Rothstein et al.[36]	12	Symptomatic	N/A	Passive	+	+	–	–	–	2	Same day	12	Expert
Zwerus[7]	30	Healthy	18–79	Both	+	+	+	+	–	2	Week	2	Expert
Total	376				9	8	5	5	1	25		86	

CA: Carrying Angle.

Table 3. Inter and intrarater reliability of included studies.

	Intrarater reliability								Interrater reliability							
	Flexion		Extension		Pronation		Supination		Flexion		Extension		Pronation		Supination	
Study	ICC	SEM	ICC	SEM	ICC	SEM	ICC	SEM	ICC	SEM	ICC	SEM	ICC	SEM	ICC	SEM
Armstrong et al.[21]																
Expert (n = 3)	0.55–0.98	–	0.45–0.98	–	0.96–0.99	–	0.96–0.99	–	0.58–0.62	–	0.58–0.87	–	0.83–0.86	–	0.91–0.93	–
Non-expert (n = 2)	0.59–0.79	–	0.97–0.98	–	0.97	–	0.97–0.98	–								
Blonna et al.[15]																
Expert (n = 3)	–	–	–	–	–	–	–	–	0.94–0.98	–	0.94–0.98			–	–	–
Non-expert (n = 1)	–	–	–	–	–	–	–	–	0.81–0.86	–	0.76–0.78			–	–	–
Chapleau et al.[6]	0.95	–	0.97	–	–	–	–	–	–	–	–	–	–	–	–	–
Cimatti et al.[30]																
Injured	–	–	–	–	–	–	–	–	–	–	–	–	0.94	–	0.97	–
Non-injured	–	–	–	–	–	–	–	–	–	–	–	–	0.92	–	0.95	–
Fieseler et al.[3]	0.79–0.96	–	0.80–0.88	–	–	–	–	–	–	–	–	–	–	–	–	–
Flowers et al.[31]	–	–	–	–	–	7.0	–	3.7	–	–	–	–	0.79	–	0.95	–
Gajdosik[32]	–	–	–	–	0.81–0.97	–	0.81–0.97	–	–	–	–	–	–	–	–	–
Goodwin et al.[33]	0.61–0.92	–	–	–	–	–	–	–	0.56–0.91	–	–	–	–	–	–	–
Low[34]	–	–	–	–	–	–	–	–	–	5	–	–	–	–	–	–
Petherick et al.[35]	–	–	–	–	–	–	–	–	0.53	–	0.53	–	–	–	–	–
Rothstein et al.[36]	0.94–0.97	–	0.86–0.99	–	–	–	–	–	0.89–0.96	–	0.93–0.96	–	–	–	–	–
Zwerus[7]																
Active	0.76	3	0.92	2	0.90	3	0.91	3	0.86	2	0.89	1	0.92	3	0.87	3
Passive	0.74	3	0.95	2	0.86	3	0.90	3	0.79	2	0.85	2	0.91	3	0.82	3

ICC: intraclass correlation coefficient; SEM: standard error of measurement.

heterogeneity, clinically the studies were also very heterogeneous. Therefore, we decided to review the ICCs narratively. The intrarater and interrater reliability of included studies is summarized in Table 3.

Validity

One study investigated the validity of goniometric elbow measurements.[6] They compared the goniometric measurements of the elbow by one examiner with radiographic measurements of 51 healthy volunteers by two examiners. They found maximal errors of the goniometric measurements of 10.3° for extension, 7.0° for flexion, 11.5° for total ROM and 6.5° for the carrying angle.

Intrarater reliability

Six studies investigated the intrarater reliability for flexion,[3,6,7,21,33,36] showing fair to excellent reliability. Results for expert and non-expert raters[21] and passive and active measurements[7] were similar. Zwerus et al.[7] calculated a standard error of measurement (SEM) of 3°.

Five studies investigated the intrarater reliability for extension,[3,6,7,21,36] showing fair to excellent reliability. One study calculated a SEM of 2°.[7]

Four studies reported the intrarater ICCs for pronation and supination, all showing excellent reliability.[7,21,31,33] Two studies used the SEM, showing 3° and 7°, respectively, for pronation and 3° and 4° for supination.[7,31] One study investigated the ICC in measurements of the carrying angle, showing excellent reliability.[6]

Interrater reliability

Six studies investigated the interrater reliability for flexion, showing fair to excellent reliability. Results for expert and non-expert raters[29] and passive and active measurements[7] were similar. Two studies reported a SEM of 2° and 5°, respectively.[7,34]

Five studies tested the interrater reliability in extension, showing fair to excellent reliability.[7,21,29,35,36] Results for expert and non-expert raters[29] and passive and active measurements[7] were similar. Zwerus et al.[7] reported a SEM of 2°.

Five studies investigated the interrater reliability of pronation and supination,[7,21,30–32] all showing excellent reliability. Results for injured and non-injured subjects[30] and passive and active measurements[7] were similar. Zwerus et al.[7] reported a SEM of 3° for both pronation and supination.

Discussion

The reliability and validity of the UG in measurements of the elbow was systematically examined. Based on 12 included studies, the overall reliability of the UG ranged among studies, from poor to excellent. There was no clear difference between intra- and interrater reliability. The most striking outlier included deviating measurements of one expert rater for inter- and intrarater reliability for flexion and extension in the study by Armstrong et al.[21] without providing a clear explanation.

The reliability for flexion, extension, pronation and supination was similar. The hypothesis that the reliability of the UG is higher in the hands of an expert examiner seems partially true.[6] Amstrong et al.[21] did not find a difference in intrarater reliability in expert versus non-expert examiners, but they did give all examiners specific directions about arm positions and goniometric alignment. In the study from Blonna et al.,[29] the examiners were free to use any bony landmarks they preferred. They found a lower interrater reliability in non-expert examiners compared to expert examiners. This suggests that the reliability of the UG in non-expert examiners can be easily increased by clear instructions on goniometric alignment.

Previous literature stated that the reliability of the goniometer is higher when the same bony landmarks are used.[22] In this systematic review the studies using bony landmarks may not show a higher reliability. It is important to mention though that three out of four studies using the bony landmarks were of moderate quality.

Only one of the included studies in our systematic literature review investigated the validity of the goniometer.[6] They used radiography as reference test for goniometric measurements and found a potential maximum error of 11.5%. When precise values of ROM of the elbow are needed, they advised radiographic measurements.

Several previous studies investigated the reliability of the UG in measurements of joints other than the elbow. For example, in a study by Brosseau et al.[40] an excellent intrarater reliability of the UG for knee flexion was found. They also stated that a difference of more than 5.5° in knee flexion is necessary to determine progression/change in the ROM. Kim and Kim[41] investigated the reliability of the UG in hip and shoulder measurement and found high test–retest reliability even in unskilled examiners.

In this digital era it is important to realize that a lot of research has been performed comparing the UG with other devises and methods such as an internet goniometer, a digital goniometer and VDO clip-based goniometry.[15–20,42,43] To maximize homogeneity this review focused on the UG. A future systematic review can be performed including and comparing these devises. It might be interesting to compare these devises and measuring methods with radiographic measurements to objectify their validity.

In all systematic reviews, there is a risk of overlooking papers. To minimize this risk an extensive search with sensitive search criteria and synonyms was performed, in collaboration with an experienced librarian. Also the included papers were scanned for other suitable studies.

Another limitation is the diversity and heterogeneity of the included articles. To avoid this clinical heterogeneity strict inclusion criteria were applied. Nonetheless, there was a high diversity in study methods, such as blinding or not blinding the examiners from their own or other measurements. There was a high difference in interval of measurements, which can influence the outcomes. Furthermore, four articles did not clarify whether the examiners were expert or non-expert examiners, which makes it more difficult to interpret the outcomes. Finally only five studies were of high quality; the other seven studies were of moderate of low quality.

The strength of this article is that it gives a clear overview of the research performed and their outcomes.

Conclusions drawn from this literature review are also limited because of the use of ICCs to assess reliability. It would be favourable to use a different approach to assess the agreement between two quantitative methods of measurement, because it possibly draws a misleading conclusion and is hard to transfer to an individual patient.

The ICC is in the general literature defined as a ratio of variance of interest over total variance (composed of variance of interest and error variance). In reliability studies for ROM measurement, the variance among patients is often considered as the variance of interest.[44] Because the ICC uses variance between subjects' ROM measurements to calculate reliability, a large variation between subjects will lead to a higher ICC, even though the measurement error is similar.[45] This could possibly draw a misleading conclusion of good reliability. For example, this could have been the case with the higher ICCs for measurements in injured subjects compared to non-injured subjects and/or the higher ICCs in non-experts compared to expert examiners.[21,30] Furthermore, ICCs are not presented as metric units and can therefore not be directly applied on an individual.

There are several ways to avoid the aforementioned problems induced by the use of ICC. Some authors already made efforts to use other ways to assess the reliability such as the SD, SEM and SDD.[7,31,34] Contribution of variance caused by subjects ($Var_{subject}$), occasion ($Var_{occasion}$) or measurement error (Var_{error}) can be determined using variance components analysis, in order to calculate the SEM and the smallest detectable difference (SDD). SEM can be calculated using the following formula: $SEM = \sqrt{(Var_{occasion} + Var_{error})}$ and SDD using the following formula: $SDD = \sqrt{2} * 1.96 * \sqrt{(Var_{occasion} + Var_{error})}$. These measurements focus on the variance of different sources of error instead of the ratio of variances (ICC) and are presented in the metric unit of the measurements (degrees, in our case), which makes it easier to interpret for the use in clinical practice.[44,45]

Bland and Altman (B&A) proposed an alternative analysis, based on the mean difference and limits of agreement.[37] B&A plot analysis evaluates a bias between mean differences and estimates an agreement interval in which 95% of the differences between two measurements fall. Based on this plot (presented in a certain unit or percentage), the clinician can decide whether the limits are acceptable or not. Therefore, we suggest the use of SEM/SDD supplemented with B&A analysis for future research on the reliability of goniometric measurements.

Conclusion

Twelve studies reported on the reliability of the UG in measurements of the elbow were included. Overall, the studies showed at least a fair intra- and interrater reliability of the UG. The reliability of the UG in non-expert examiners can be increased by clear instructions on goniometric alignment. For future research, it would be favourable to use another statistical approach to substitute or supplement to ICCs.

Acknowledgements

The authors would like to thank Wichor Bramer, librarian at the Erasmus Medical Centre for his assistance in the systematic search of the literature.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Ethical Review and Patient Consent

No human participants were used for this study, therefore ethical approval was not required.

ORCID iD

Elisa L Zwerus http://orcid.org/0000-0003-4656-6053

References

1. Soucie JM, Wang C, Forsyth A, et al. Range of motion measurements: reference values and a database for comparison studies. *Haemophilia* 2011; 17: 500–507.

2. Wright RW, Steger-May K, Wasserlauf BL, et al. Elbow range of motion in professional baseball pitchers. *Am J Sports Med* 2006; 34: 190–193.
3. Fieseler G, Molitor T, Irlenbusch L, et al. Intrarater reliability of goniometry and hand-held dynamometry for shoulder and elbow examinations in female team handball athletes and asymptomatic volunteers. *Arch Orthop Trauma Surg* 2015; 135: 1719–1726.
4. Gunal I, Kose N, Erdogan O, et al. Normal range of motion of the joints of the upper extremity in male subjects, with special reference to side. *J Bone Joint Surg Am* 1996; 78: 1401–1404.
5. Karagiannopoulos C, Sitler M and Michlovitz S. Reliability of 2 functional goniometric methods for measuring forearm pronation and supination active range of motion. *J Orthop Sports Phys Ther* 2003; 33: 523–531.
6. Chapleau J, Canet F, Petit Y, et al. Validity of goniometric elbow measurements: comparative study with a radiographic method. *Clin Orthop Relat Res* 2011; 469: 3134–3140.
7. Zwerus EL, Willigenburg NW, Scholtes VAB, et al. Normative values and affecting factors for the elbow range of motion. *Shoulder & Elbow*. Epub ahead of print 11 September 2017. DOI: 10.1177/1758573217728711.
8. Jobe FW and Nuber G. Throwing injuries of the elbow. *Clin Sports Med* 1986; 5: 621–636.
9. King JW, Brelsford HJ and Tullos HS. Analysis of the pitching arm of the professional baseball pitcher. *Clin Orthop Relat Res* 1969; 67: 116–123.
10. Docherty MA, Schwab RA and Ma OJ. Can elbow extension be used as a test of clinically significant injury? *South Med J* 2002; 95: 539–541.
11. Lennon RI, Riyat MS, Hilliam R, et al. Can a normal range of elbow movement predict a normal elbow x ray? *Emerg Med J* 2007; 24: 86–88.
12. Morrey BF, Askew LJ and Chao EY. A biomechanical study of normal functional elbow motion. *J Bone Joint Surg Am* 1981; 63: 872–877.
13. Sardelli M, Tashjian RZ and MacWilliams BA. Functional elbow range of motion for contemporary tasks. *J Bone Joint Surg Am* 2011; 93: 471–477.
14. Raiss P, Rettig O, Wolf S, et al. [Range of motion of shoulder and elbow in activities of daily life in 3D motion analysis]. *Z Orthop Unfall* 2007; 145: 493–498.
15. Blonna D, Zarkadas PC, Fitzsimmons JS, et al. Validation of a photography-based goniometry method for measuring joint range of motion. *J Shoulder Elbow Surg* 2012; 21: 29–35.
16. Ferriero G, Vercelli S, Sartorio F, et al. Reliability of a smartphone-based goniometer for knee joint goniometry. *Int J Rehabil Res* 2013; 36: 146–151.
17. Milanese S, Gordon S, Buettner P, et al. Reliability and concurrent validity of knee angle measurement: smart phone app versus universal goniometer used by experienced and novice clinicians. *Man Ther* 2014; 19: 569–574.
18. Naylor JM, Ko V, Adie S, et al. Validity and reliability of using photography for measuring knee range of motion: a methodological study. *BMC Musculoskelet Disord* 2011; 12: 77.
19. Ockendon M and Gilbert RE. Validation of a novel smartphone accelerometer-based knee goniometer. *J Knee Surg* 2012; 25: 341–345.
20. Werner BC, Holzgrefe RE, Griffin JW, et al. Validation of an innovative method of shoulder range-of-motion measurement using a smartphone clinometer application. *J Shoulder Elbow Surg* 2014; 23: e275–e282.
21. Armstrong AD, MacDermid JC, Chinchalkar S, et al. Reliability of range-of-motion measurement in the elbow and forearm. *J Shoulder Elbow Surg* 1998; 7: 573–580.
22. Fish DR and Wingate L. Sources of goniometric error at the elbow. *Phys Ther* 1985; 65: 1666–1670.
23. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Open Med* 2009; 3: e123–e130.
24. Lucas N, Macaskill P, Irwig L, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC Med Res Methodol* 2013; 13: 111.
25. Gorgos KS, Wasylyk NT, Van Lunen BL, et al. Inter-clinician and intra-clinician reliability of force application during joint mobilization: a systematic review. *Man Ther* 2014; 19: 90–96.
26. May S, Chance-Larsen K, Littlewood C, et al. Reliability of physical examination tests used in the assessment of patients with shoulder problems: a systematic review. *Physiotherapy* 2010; 96: 179–190.
27. Barrett E, McCreesh K and Lewis J. Reliability and validity of non-radiographic methods of thoracic kyphosis measurement: a systematic review. *Man Ther* 2014; 19: 10–17.
28. Cicchetti DV. Multiple comparison methods: establishing guidelines for their valid application in neuropsychological research. *J Clin Exp Neuropsychol* 1994; 16: 155–161.
29. Blonna D, Zarkadas PC, Fitzsimmons JS, et al. Accuracy and inter-observer reliability of visual estimation compared to clinical goniometry of the elbow. *Knee Surg Sports Traumatol Arthrosc* 2012; 20: 1378–1385.
30. Cimatti B, Marcolino AM, Barbosa RI, et al. A study to compare two goniometric methods for measuring active pronation and supination range of motion. *Hand Ther* 2013; 18: 57–63.
31. Flowers KR, Stephens-Chisar J, LaStayo P, et al. Intrarater reliability of a new method and instrumentation for measuring passive supination and pronation: a preliminary study. *J Hand Ther* 2001; 14: 30–35.
32. Gajdosik RL. Comparison and reliability of three goniometric methods for measuring forearm supination and pronation. *Percept Mot Skills* 2001; 93: 353–355.
33. Goodwin J, Clark C, Deakes J, et al. Clinical methods of goniometry: a comparative study. *Disabil Rehabil* 1992; 14: 10–15.
34. Low JL. The reliability of joint measurement. *Physiotherapy* 1976; 62: 227–229.
35. Petherick M, Rheault W, Kimble S, et al. Concurrent validity and intertester reliability of universal and fluid-

based goniometers for active elbow range of motion. *Phys Ther* 1988; 68: 966–969.
36. Rothstein JM, Miller PJ and Roettger RF. Goniometric reliability in a clinical setting. Elbow and knee measurements. *Phys Ther* 1983; 63: 1611–1615.
37. Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307–310.
38. Bland JM. *Introduction to medical statistics*. Oxford: Oxford University Press, 2006.
39. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ* 2003; 327: 557–560.
40. Brosseau L, Tousignant M, Budd J, et al. Intratester and intertester reliability and criterion validity of the parallelogram and universal goniometers for active knee flexion in healthy subjects. *Physiother Res Int* 1997; 2: 150–166.
41. Kim SG and Kim EK. Test-retest reliability of an active range of motion test for the shoulder and hip joints by unskilled examiners using a manual goniometer. *J Phys Ther Sci* 2016; 28: 722–724.
42. Hoffmann T, Russell T and Cooke H. Remote measurement via the Internet of upper limb range of motion in people who have had a stroke. *J Telemed Telecare* 2007; 13: 401–405.
43. Meislin MA, Wagner ER and Shin AY. A comparison of elbow range of motion measurements: smartphone-based digital photography versus goniometric measurements. *J Hand Surg Am* 2016; 41: 510–515e1.
44. Roebroek ME, Harlaar J and Lankhorst GJ. The application of generalization theory to reliability assessment: an illustration using isometric force measurements. *Phys Ther* 1993; 73: 386–395.
45. Scholtes VA, Terwee CB and Poolman RW. What makes a measurement instrument valid and reliable? *Injury* 2011; 42: 236–240.