J Shoulder Elbow Surg (2013) 22, 1380-1394

ELSEVIER

Journal of Shoulder and Elbow Surgery

www.elsevier.com/locate/ymse

ELBOW AND BASIC SCIENCE

Elbow-specific clinical rating systems: extent of established validity, reliability, and responsiveness

Bertram The, MD, PhD[a,*], Inge H.F. Reininga, PhD[b], Mostafa El Moumni, MD[b], Denise Eygendaal, MD, PhD[c]

[a]*Department of Orthopaedic Surgery, Sir Charles Gairdner Hospital, Perth, WA, Australia*
[b]*Department of Trauma Surgery, University Medical Center Groningen, Groningen, The Netherlands*
[c]*Department of Orthopaedic Surgery, Amphia Hospital, Breda, The Netherlands*

Background: The modern standard of evaluating treatment results includes the use of rating systems. Elbow-specific rating systems are frequently used in studies aiming at elbow-specific pathology. However, proper validation studies seem to be relatively sparse. In addition, these scoring systems might not always be used for appropriate populations of interest. Both of these issues might give rise to invalid conclusions being reported in the literature. Our aim was to investigate the extent to which the available elbow-specific outcome measurement tools have been validated and the quality of the validation itself. We also aimed to provide characteristics of the populations used for validation of these scales to enable clinicians to use them appropriately.

Methods: A literature search identified 17 studies of 12 different elbow-specific scoring systems. These were assessed for validity, reliability, and responsiveness characteristics. The quality of these assessments was rated according to the Consensus Based Standards for the Selection of Health Measurement Instruments (COSMIN) checklist criteria, a standardized and validated tool developed specifically for this purpose.

Results: Currently, the only elbow-specific rating system that is validated using high-quality methodology is the Oxford Elbow Score, a patient-administered outcome measure tool that has been validated on heterogeneous study populations.

Conclusion: Other rating systems still have to be proven in the future to be as good as the Oxford Elbow Score for clinical or research purposes. Additional validation studies are needed.

Level of evidence: Basic Science, Validation of Outcome Instruments, Systematic Review.

Keywords: Clinical rating systems; elbow; validity; reliability; responsiveness; population

The current standard of treatment evaluation includes the use of clinical rating systems. These rating scales assess multiple areas of interest (such as pain, range of motion, etc) and yield a summarized value that is supposed to correspond to the effect of the diseased joint on the patient. These scoring systems can be designed to be administered to patients[4,5,20,21] or to the physician.[2,15-19] Initially, the latter was mostly the focus of these scoring systems, but more recently, several scoring systems, based purely on the patient's perception of the functioning of the joint, have been developed. These different perspectives used for

Investigational Review Approval was not required for this study.

*Reprint requests: Bertram The, MD, PhD, 2/130 Keightley Rd, Shenton Park, WA 6008, Perth, Australia.

E-mail address: bertramthe@hotmail.com (B. The).

1058-2746/$ - see front matter © 2013 Journal of Shoulder and Elbow Surgery Board of Trustees.
http://dx.doi.org/10.1016/j.jse.2013.04.013

judging an outcome of a certain intervention do not necessarily yield similar results and should be considered to be complementing parts of a larger spectrum.

Although many systems have been developed and extensively validated to assess treatment outcome for surgery on the lower extremity, this is true to a much lesser extent for the upper extremity and is certainly not true when considering elbow-specific rating systems. Using elbow-specific rating systems seems to be a sensible choice when handling conditions limited to the elbow. This can be expected to result in achieving better validity and responsiveness characteristics at least, and perhaps also higher reliability scores. However, high-quality validation is required to ensure the validity of the results.

The validity, reliability, and responsiveness characteristics of a measurement tool are not solely inherent to the rating system itself but are more or less a product of the rating system properties and the characteristics of the population it is being used for. This is why it should be borne in mind that established validity characteristics might not be applicable when using the rating system for a different population. This means that an already validated outcome measurement tool might have to be validated more than once to justify its use in different populations. Although this perspective may be generally accepted, it is often neglected, and rating systems are used for pathophysiologic and surgical conditions or interventions for which they have never been properly validated. This may lead to incorrect conclusions drawn from studies that have used such rating systems as a primary outcome.

Our aim is to investigate the extent to which the available elbow-specific outcome measurement tools have been validated and the quality of the validation itself. We also provide characteristics of the populations used for validation of these scales to enable clinicians to use them appropriately.

Materials and methods

Search strategy for identifying rating systems

Two authors (B.T. and I.R.) conducted a PubMed search using the MeSH terms “Outcome Assessment (Health Care)” *and* “Elbow” (yielding 270 reports). We also conducted conventional PubMed searches using the keywords “elbow scoring system” (132 reports), “elbow rating system” (70 reports), and “elbow validation” (156 reports).

The resulting abstracts were retrieved and articles were selected if deemed relevant to our topic; that is, when the main focus was an elbow-specific rating system (development, validation, or assessment of clinimetric properties) or an overview of elbow rating systems. Only reports of the English versions of the rating systems were selected. This resulted in a selection of 9 articles. These articles were examined, including a hand search of the reference lists, which resulted in the selection of an additional 8 reports, bringing the total to 17 articles.[1-12,15-19,21]

We scrutinized each report for assessment of the following features:

Validity

Validity refers to the property of a method to measure what it aims to measure. Perhaps the most common way is to assess *criterion validity.* This is performed by using a gold standard to assess the validity of a newly developed measurement instrument. For example, intraoperative findings can serve as a gold standard to assess the ability of a new imaging technique to detect cartilage lesions.

However, because rating systems often aim to assess entities for which a gold standard is not easy to define (for example, pain), they are mostly assessed by judging content validity and construct validity. *Content validity* is a qualitative judgement of experts who deem the items of a score relevant or not to what the score aims to measure. *Construct validity* is assessed by quantification of the relationship between the measurement tool and a theoretic concept (construct). For example, because paralysis of a limb certainly leads to a degree of disability, a rating system aimed at quantifying disability is expected to yield different degrees of disability scores with different degrees of paralysis.

The most widely used approach is to investigate the strength of association between the new measurement tool and an established, validated system that overlaps to a certain extent in what it aims to measure, also known as *convergent validity.* Another less frequently applied approach is to establish the lack of association between the new rating system and established rating systems that aim to measure a completely unrelated concept. This is known as *divergent* or *discriminant validity.*

Reliability

Reliability refers to the reproducibility of the rating system results: Does the scoring system yield the same result when given to the patient on 2 different days (intraobserver reliability)? Does the scoring system yield the same results when given to 2 different clinicians (interobserver reliability)? For patient-administered outcomes only *intraobserver reliability* is an issue, which can be tested by administering the same survey twice to a patient within a certain interval in which no real change is assumed to have occurred. For physician-administered outcome measures, *inter observer reliability* can be assessed additionally. This refers to the agreement of obtained scores when the same patient is rated by more than 1 physician at the same moment.

Another measurement property, which appertains specifically to multi-item questionnaires, is *internal consistency.* This is a measure based on the inter-relatedness between different items on the same subscale (eg, different items that all are supposed to reflect impairment in function of the elbow) on a larger test. It measures whether several items that propose to measure the same general construct (eg, elbow function) produce similar scores and is mostly measured using Cronbach α. High internal consistency of a subscale is desirable, although very high inter-relatedness of the different items (eg, a Cronbach α exceeding 0.95) might be a sign of redundancy: they are apparently similar to such an extent that you might just need to score one of those items to obtain the same information.