Open Journal of Orthopedics, 2015, 5, 264-276
Published Online September 2015 in SciRes. http://www.scirp.org/journal/ojo
http://dx.doi.org/10.4236/ojo.2015.59036

What Is the Accuracy of Shoulder Range of Motion Measurements on Physical Exam?

Yousef Shishani, Janice Flocken, Reuben Gobezie

Shoulder and Elbow Surgery, The Cleveland Shoulder Institute, Beachwood, USA
Email: Reuben.Gobezie@UHHospitals.org

Received 2 August 2015; accepted 29 August 2015; published 1 September 2015

Abstract

The purpose of this study is to investigate a new method for measuring shoulder range of motion (ROM) in an orthopedic practice utilizing a smartphone application to improve accuracy from physical exam typically used in research. Our aim is to evaluate the application, Physio2Go (P2G), which uses a virtual goniometer, assessing validity by comparing its measurements to those taken by a universal goniometer (UG). Two observers of varying clinical experience, a research assistant and research fellow, compared measurements. Statistically, we used the intra-class correlation coefficient (ICC), standard error of measurement (SEM), and the Pearson correlation coefficient (PCC). Following validation we tested P2G in symptomatic postoperative shoulder patients measuring forward flexion (FF) and external rotation (ER). We compared P2G measurements to visual estimation (VE) done by a fellowship trained orthopedic surgeon. Statistically we used ICC, Bland-Altman plots with 95% limits of agreement (LOA), and scatter plots. We examined the impact of the application using Welch's t-test comparing pre-to-postoperative ROM improvements using the values obtained by P2G and VE. We found high intra-rater reliability of P2G for both observers, substantial correlation between UG and P2G measurements, highly correlated inter-observer reliability for UG and P2G, and statistically significant PCC values ($p < 0.05$). As expected, ROM measurements of symptomatic patients comparing P2G and VE measurements demonstrated lower correlation. Bland-Altman plots demonstrated wide confidence intervals; scatterplots and histograms confirmed low agreement among measurement methods. Clinical application demonstrated varying statistical significance depending on whether measurements were done by P2G or VE. Our study found that P2G provided superior reliability compared to the customary physical exam routinely used for orthopedic research. The value of using this application instead of a UG is the ease of use and the ability for any member of the healthcare team, regardless of clinical experience to be able to produce reliable and valid measurements.

Keywords

Shoulder, Range of Motion, Digital Assessment, Goniometry, Smartphone Application

How to cite this paper: Shishani, Y., Flocken, J. and Gobezie, R. (2015) What Is the Accuracy of Shoulder Range of Motion Measurements on Physical Exam? *Open Journal of Orthopedics*, **5**, 264-276. http://dx.doi.org/10.4236/ojo.2015.59036

1. Introduction

Accurate measurement of the shoulder joint range of motion (ROM) is imperative in assessing postoperative clinical outcomes. Many research papers designed to evaluate clinical interventions utilize differences in ROM obtained by visual inspection to draw conclusions on differences in outcomes. The validity of these conclusions is based on the accuracy of these measurements. However, the literature denotes that visual inspection for ROM measurements results in inaccuracy based on intraobserver variability and reproducibility [1]-[3]. Orthopedic investigators are in need of quick and cost-effective methods of obtaining accurate measurement for ROM on physical exam. The purpose of this study is to investigate a new method for measuring shoulder ROM in an orthopedic practice utilizing a smartphone application to improve accuracy from physical exam typically used for outcomes in research.

Goniometry has long been considered the golden standard when measuring a joint ROM [4]. Traditionally, the universal goniometer (UG) has been used to perform these measures clinically due to its relatively inexpensive cost, strong intra-class correlation coefficient (ICC), and high reliability of measurements, even when different clinicians are performing the measurements [5] [6]. Techniques used to measure the wide ROM of the shoulder joint need to be consistent and standardized in order to provide reliable and reproducible results [7].

Despite its advantages, goniometry can also be time-consuming and demands proper training. As patient volumes are constantly rising, the time that the surgeon can allocate in an office visit has to be properly managed; thus surgeons are relying more heavily on medical assistants, physician assistants and research teams to help with in-office assessments. It is necessary that these teams adopt reliable and time-efficient methods that can be learned quickly and performed accurately without extensive clinical experience.

In the past decade there have been numerous studies investigating the accuracy, reliability, and ease of use for newer ROM measurement methods such as digital photography and inclinometers [1] [8]. Digital and high speed photography has shown excellent ICC and reliability for internal rotation (IR), external rotation (ER), flexion (FF), and abduction measurements [5] [8] [9]. Digital photography also allows continuity between clinical examination and patients' follow-ups through physical therapy [2]. However, digital photography may not be the most practical for immediate in-clinic results as they must first be uploaded to an external computer and then digitally measured.

Similarly, use of the hand-held digital inclinometer has garnered clinical support for its usefulness and accuracy in measuring ROM as well as strength. These devices rely on gravity to record the change in motion on a 360˚ scale [5] [10]. Unfortunately the cost and training required to use these devices can make them impractical for medical practices that lack the necessary resources. What we are beginning to see now is a shift from actual, expensive and measuring tools to smartphone applications (apps) that mimic these tools without the cost.

In 2013 more than 80% of clinicians used smartphones in their professional activities, and more than 50% of clinicians used tablets. These numbers are expected to continue to rise [11]. Numerous phone or tablet applications are now available for clinical use for free or a very small fee. Those designed to measure ROM have proven reliable and accurate [7] [12]-[15]. Some of these apps use built-in sensors such as a magnetometer or gyrosensor to determine ROM relative to the phone's position in space or a change in angle as the limb moves [7] [13] [15]. These apps function as inclinometers, for a much lower price than a traditional inclinometer and most show excellent reliability for healthy subject ROM measures when compared to UG measures [13]-[15]. The few studies that have measured ROM in symptomatic shoulder patients show slightly more errors in measurement, but still demonstrate substantial correlation with the UG measurements [7] [15].

However, smartphone based inclinometers raise many concerns for their applicability in actual clinics. The use of these applications requires the device to be held by the patient or attached to the patient's body. As many patients, especially older ones, don't have the strength to hold a phone or tablet, they won't be able to use these applications. If the device is attached to the patient's wrist, one encounters the risk of the patient "cheating" the second axis by bending at the elbow to extend their ROM [7]. Also, these applications work off the gravitational directions of the earth's plane, which are not the same planes as the body used in ROM. While the supine position can align the horizontal axis with the body position, this is not possible for all directions necessary for shoulder ROM, which use the midline of the thorax instead of the true vertical [16], nor is it practical for clinical use.

A different type of smartphone application used to measure ROM is the digital goniometer. These applications allow the user to take a photograph of the patient and digitally measure the ROM immediately on the photograph, providing a quick measurement that allows the user to account for bony landmarks and patient compensations. To our knowledge, only one such application has been evaluated, DrGoniometer; this application has shown high intra- and inter-rater correlation values and reliable agreement when compared to a UG when measuring elbow joint ROM [12]. To our knowledge, no studies have been conducted using a smartphone virtual goniometer app to test ROM of the shoulder joint. Our aim is to evaluate ROM using the smartphone application, Physio2Go (P2G), which uses a virtual goniometer, available through iTunes stores to all iOS operating system devices: iPad, iPhone etc.

By using P2G we will compare the digital shoulder ROM measurements with the visual estimation (VE) routinely done during physical exams. We aim to show that when taken in an actual clinical setting VE is inaccurate and varies widely, well over the ±10˚ acceptable for clinical use [17]. Isolating a particular cohort of our patients undergoing reverse total shoulder arthroplasty, we aim to demonstrate the advantage of using P2G instead of VE when applied to specific patient pre-to-postoperative comparisons. We believe that few practices today use physical goniometers for each patient exam; it can be clumsy, take too much time, or be inaccurate if the observer isn't well trained. By using an easy smartphone application, this will ideally allow clinicians to measure ROM quickly and reliably, keep accurate records, and allow a cohesive congruity to the electronic medical records required in today's healthcare systems.

2. Methods

2.1. Subjects

Following Institutional Review Board approval we began to enroll healthy subjects for validation. After giving consent for participation, six healthy female (11 shoulders) subjects, average age 30 years (23 - 34) were used in this part of the study. Subjects were included if they had no self-reported history of previous shoulder injury, exclusion criteria was previous or current shoulder injury. Both shoulders were used for measurement, regardless of dominant arm. Although we attempted to also enroll males for validation, this was done on one day due to staff availability and only females consented to participate.

Following the initial validation (validation method detailed in 2.4a), 97 symptomatic patients (47 men) with a mean age of 59.8 (18 - 88) were selected at random, and consented for evaluation as part of their standard postsurgical follow up. Patients included were those who had undergone shoulder arthroscopy procedures (n = 41) within 6 months or shoulder arthroplasty procedures within 12 months (n = 56). Patients who had serious postoperative complications such as fractures were excluded along with any revision cases. Patients who had obviously varying ROM (greater than 10˚ by visual estimation) upon repeated testing were also excluded.A sample size of 97 was chosen from a significance of 0.05 (alpha) and power of 0.20 (beta) considering 95% confidence intervals with a 5% margin of error. Validation was done in November 2014; symptomatic evaluation was completed between November 2014 and June 2015.

2.2. Observers

To evaluate inter- and intra-observer reliability a research assistant with 1 year clinical experience and a research fellow with 10 years of experience performed all goniometric and tablet application measurements. (Calculations used to calculate intra-rater reliability described in section 2.5). Once interobserver reliability was established the research assistant performed all tablet application measurements. VE was performed by a fellowship trained orthopedic surgeon with over 10 years practice in shoulder and elbow surgery.

2.3. Devices

Shoulder ROM was measured using three methods: a standard double-arm plastic universal goniometer (UG) (Clinoril$^{®}$), where one face was covered to prevent examiners from reading the measurement; the smartphone application, Physio2Go, (Gerard Vehof Physiotherapy, 2015) which was downloaded onto two Ipad-mini (iOS 8.3); and VE performed by a fellowship trained orthopedic surgeon.

2.4. Procedure

2.4.1. Validation

Prior to the start of validation the research assistant and research fellow had two weeks to familiarize themselves with P2G and practice using the app. The surgeon was shown the application and how it would be used to measure ROM but did not use the application. Both research assistant and research fellow also re-familiarized themselves with the process of using the UG to measure FF and practiced identifying the relevant bony landmarks of the shoulder.

Healthy volunteers stood perpendicular to the wall of exam room at a position marked by tape on the floor. They were instructed to forward flex the arm farther from the wall as high as they could until end of motion. The research assistant and research fellow each measured FF using the UG twice, measurement was read by an independent observer and recorded. Patient then repeated motion for measurement using P2G (**Figure 1**). Observer stood directly in line with shoulder at a standard distance marked by tape on the floor and used P2G to take a picture of the patient's FF (**Figure 2**). Each observer used P2G to measure ROM twice. The screen of the application was taped over to prevent the observer from knowing what the first measurement read. The entire measuring protocol was repeated for the other shoulder with the patient turning 180˚ so the first measured shoulder was now closer to the wall.

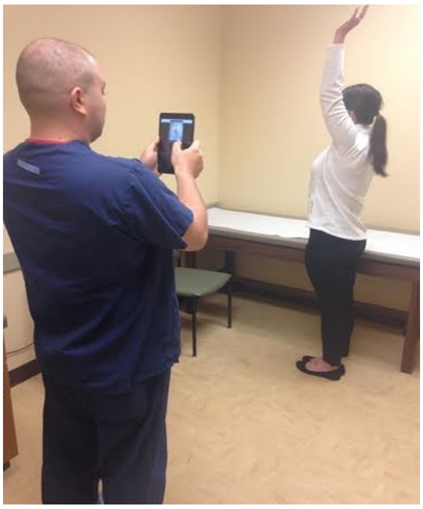

Figure 1. Setup of validation, research fellow measuring forward flexion in healthy subject.

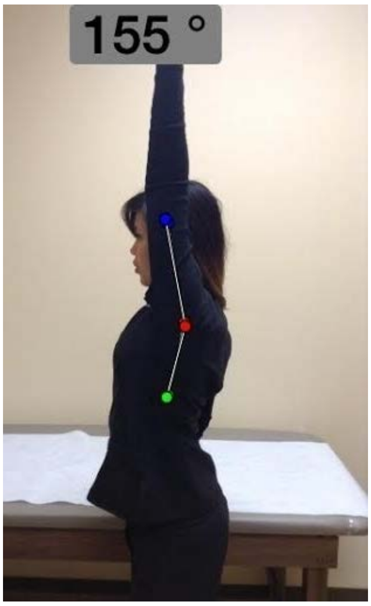

Figure 2. Physio2Go measurement of forward flexion in healthy subject.

2.4.2. Symptomatic Patients

After patients consented to participation the research assistant measured FF and ER using the smartphone application Physio2Go at a follow up visit (**Figure 3** and **Figure 4**). The research assistant then left the exam room and informed the surgeon which arm needed to be measured visually. Upon leaving the patient exam room the surgeon told the research assistant his FF and ER measurements and the research assistant recorded these and the application measurements immediately. The surgeon was blinded from results of the application and was not made aware of how his estimates compared to the estimates of the application until after the close of the study.

FF was measured post-operatively as part of the patient's standard follow up protocol. Patient was instructed to stand perpendicular to exam room wall with affected shoulder closer to examiner. Examiner identified bony landmarks on patient and instructed patient to lift arm in the plane of the body as high as possible, keeping elbow extended if possible. Photograph was taken from a standard distance, marked on exam floor, as close to the level of the patient's acromion process as possible and ROM angle was measured and recorded before the examiner left the room.

For ER measurements patients sat in an armless chair, back to the wall, and were instructed to keep shoulders straight, parallel to the wall. The examiner demonstrated the motion and instructed the patient to bend their arm at the elbow to 90˚ flexion. Patients then brought the hand away from the plane of the body, keeping the elbow against their side, until they couldn't extend the arm without jeopardizing parallel shoulder placement. In-application photograph was taken from above the patient, in line with the acromion, the angle was measured from between the plane of the humerus and the elbow (**Figure 4**).

The attending orthopedic surgeon measured ROM for all patients in the cohort. As is standard practice for office visits, the physician estimated the ROM visually. Patient was instructed to lift the postoperative arm forward as high as possible to estimate FF. Patient was then instructed to hold elbow in to their side and bring hand away from the body. The surgeon reported his estimates to the research assistant upon leaving the patient room.

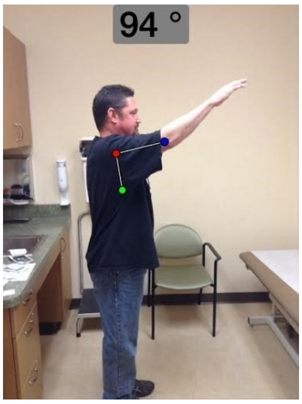

Figure 3. Forward flexion measured by Physio2Go in symptomatic patient.

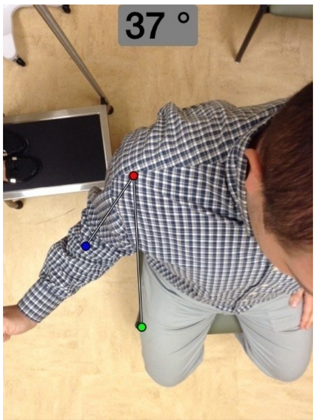

Figure 4. External rotation measured by Physio2Go in symptomatic patient.

2.5. Statistical Analysis

In order to determine validity of P2G we compared it to the UG using the intraclass correlation coefficient (ICC (2, 1)). The ICC (2, 1) is a two-way random effect model (observers and subjects are both treated as random effects) with a single measure and agreement for each measurement, hereon ICC will refer to ICC (2, 1). ICC was interpreted using the original scale determined by Landis and Koch such that: 0.00 - 0.02, slight correlation; 0.21 - 0.40, fair correlation; 0.41 - 0.60, moderate correlation; 0.61 - 0.80, substantial correlation; and 0.81 - 1.00, almost perfect correlation [18].

Intra-rater and interrater reliability was measured by the ICC for the research fellow and research assistant across two trials of FF in healthy subjects, for both UG and P2G. This was expressed as ICC with a 95% confidence interval. The standard error of measurement (SEM) was also used as a reliability analysis for each observer and each method. SEM was calculated by the equation:

$$\text{SEM} = \text{SD} * \sqrt{(1 - \text{ICC})}$$

SEM is inversely related to reliability, *i.e.* larger SEM values indicate lower SEM. Although SEM is a 68% confidence interval, it is the ubiquitous standard for ROM research and is used along with 95% confidence intervals. Pearson correlation coefficient (PCC) was calculated for an additional relationship measure of P2G compared to UG for both evaluators; higher PCC (closer to 1) indicates a stronger correlation between values.

To determine agreement and reliability between VE and P2G we used ICC with both 95% confidence intervals and SEM. We also used Bland-Altman plots to investigate systematic differences and assess 95% limits of agreement (LOA) as a second, visual indication of agreement. LOA is calculated as LOA = mean difference ± 1.96 (SD). Good agreement would be indicated by random differences around the zero difference line. Similar to SEM, lower LOA indicate that the two methods of measurement are equivalent, larger LOA indicate an ambiguous relationship between methods. Agreement between VE and P2G was also assessed by scatter plots.

In order to demonstrate clinical applications of the necessity to use a reliable, consistent method we used half of the patients who had reverse total shoulder arthroplasty (rTSA) and ran an *F*-test for two-sample variance to determine variance. A two-sample Welch's *t*-test of statistical significance was used to compare preoperative recorded ROM values and postoperative P2G measurements and also between preoperative ROM and postoperative visually estimated values.

3. Results

3.1. Validation

Intra-rater reliability measures comparing the research fellow and research assistant can be seen in **Table 1**. Both measurement methods had comparable error and each had high reliability, ICC greater than 0.80, SEM < 3.0. The research fellow and assistant had almost identical ICC values for repeated measurements using the UG, each with low SEM. The highest intra-rater reliability was achieved by the research assistant for their repeated measures using P2G, (ICC = 0.900) with the lowest SEM of 1.4˚. ICC for the research fellow using P2G was 0.837 (SEM = 1.9˚), showing similar substantial correlation as their UG measurements.

The intra-rater reliability between UG and P2G measurements was higher for the research assistant than the research fellow (**Table 2**). Both ICC values were high enough to demonstrate substantial or almost perfect cor-

Table 1. Intra-rater reliability, intraclass correlation coefficient (ICC) and 95% confidence intervals, used to determine reliability of each examiner using both UG and P2G in the forward flexion position. SEM: standard error of measurement; FF: forward flexion.

Examiner	Method	ICC	95% Confidence Interval	SEM (˚)
Research Fellow	UG FF	0.832	0.38 - 0.96	2.6
Research Assistant	UG FF	0.833	0.13 - 0.96	2.3
Research Fellow	App FF	0.837	0.50 - 0.96	1.9
Research Assistant	App FF	0.900	0.57 - 0.97	1.4
Mean		0.851		2.1

Table 2. Intra-rater comparison of universal goniometer measurement and Physio2Go application in healthy subjects, measured in forward flexion. ICC: intraclass correlation coefficient; SEM: standard error of measurement; FF: forward flexion.

Measurement	ICC	95% Confidence Interval	SEM (˚)
Research Fellow			
FF	0.785	0.08 - 0.94	2.5
Research Assistant			
FF	0.879	0.54 - 0.97	1.8
Mean	0.832		2.2

relation ICC > 0.750 [8] [18]. SEM values for intra-rater comparison across measurement modality were similar to those observed within measurement (SEM < 3.0˚).

Comparing interobserver reliability for healthy subjects in the forward flexion position showed high degree of correlation between the two evaluators (**Table 3**). Measurements taken using UG achieved the highest correlation (ICC = 0.909). Similarly, the reliability for P2G showed substantial reliability regardless of evaluator. Both measurements demonstrated low SEM (<2.0˚). The final measure of correlation, PCC, was high for both research fellow (0.679, $p < 0.05$) and research assistant (0.777, $p < 0.01$).

3.2. Symptomatic Patients and Clinical Application

Evaluation of interobserver reliability between VE and P2G can be seen in **Table 4**. There is a high degree of reliability when VE is compared to P2G for FF, (ICC = 0.879). For ER there is slightly less agreement, (ICC = 0.682). The average ICC for reliability between VE and P2G measures was substantial, at 0.781. The SEM values for both shoulder angles measured were high, 8.9˚ for FF and 10.0˚ for ER, these were the highest SEM values of any observations.

Low visual agreement for FF measures between VE and P2G measurements can be seen in **Figure 5** and **Figure 6**. Bland-Altman plots (**Figure 7** and **Figure 8**) show low agreement and high ambiguity between the measurements in both FF and ER. For FF the mean difference between measurements was 3.3˚ with a 95% LOA ± 23.9˚ (SD of 12.2˚). There was also low agreement visualized on scatter plots and histograms for ER (**Figure 9** and **Figure 10**). For ER Bland-Altman mean difference and 95% LOA was −3.1˚ ± 27.4˚ (SD of 14.0˚).

Half of the patients who underwent rTSA were chosen at random to compare their recorded preoperative ROM values with the different postoperative values. The difference between preoperative FF values (obtained with informed consent from patient chart) and the VE of postoperative FF was significant ($p = 0.007$). Comparing these same preoperative FF values and the P2G postoperative FF values the difference was much less significant ($p = 0.023$). Similarly for ER, the difference between preoperative ER and visually estimated ER tended to indicate more significance than between preoperative and P2G measured ER), although neither reached statistical significance ($p > 0.05$).

4. Discussion

The use of smartphone applications to digitally measure ROM is unique in its ability to increase reliability while simultaneously improving clinical efficacy [7] [13] [15]. Previous improvements to ROM technology, such as digital or still photography, sacrificed quickness for reliability [2], while methods commonly used for speed, such as visual estimation, sacrificed reliability [3]. We find that physical exam is both inaccurate and potentially misleading, but can be easily replaced by the valid and accurate use of a smartphone application. There exists a clear precedence to incorporate these applications into medical practice one; effectiveness is evaluated. Therefore we chose a novel approach, investigating both effectiveness and clinical importance of a previously unstudied smartphone application, to demonstrate the importance of integrating this technology into modern day medical practice.

It is without contest that the ability to measure ROM quickly and reliably is imperative for successfully documenting clinical outcomes. The results of our study indicate strong validity for the P2G application as well as an important clinical relevance. High intra- and inter-rater reliability measurements during validation, coincident with ease of use, are encouraging for this application's potential as a widely used clinical tool.

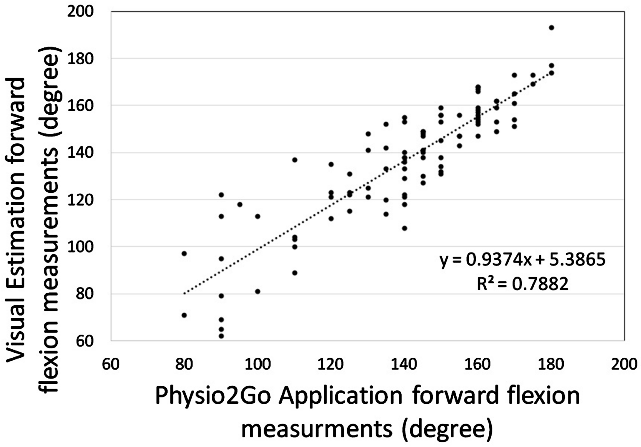

Figure 5. Scatterplot showing agreement between visual estimation and Physio2Go forward flexion measurements ($n = 98$).

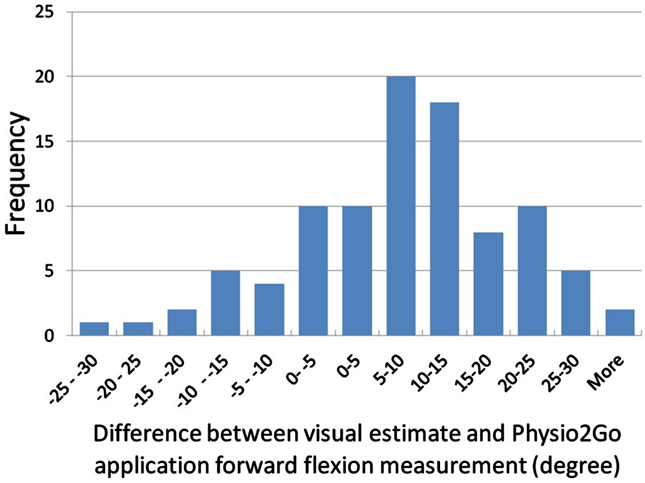

Figure 6. Histogram demonstrating differences between visual estimation and Physio2Go measurements in the forward flexion position.

Table 3. Interobserver reliability for forward flexion (FF) measurement in healthy subjects. ICC: intraclass correlation coefficient; SEM: standard error of measurement.

Measurement	ICC	95% Confidence Interval	SEM (˚)
UG FF	0.909	0.77 - 0.97	1.9
Physio2Go	0.763	0.03 - 0.94	1.8
Mean	0.836		1.85

Table 4. Comparison of visual and smartphone application measurements in symptomatic shoulder patients. ICC: intraclass correlation coefficient; SEM: standard error of measurement; LOA: limit of agreement.

Measurement	ICC	95% Confidence Interval	SEM (˚)	Mean Difference ± 95% LOA (˚)
Forward Flexion	0.879	0.82 - 0.92	8.9	3.3 ± 23.9
External Rotation	0.682	0.56 - 0.78	10.0	−3.1 ± 27.4
Mean	0.781		9.5	

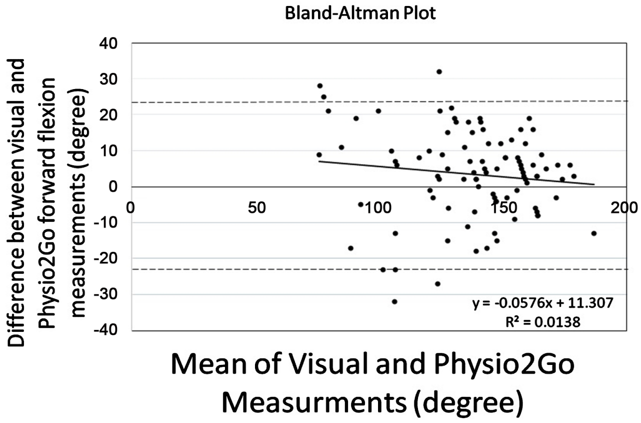

Figure 7. Bland-Altman plot of averages vs. differences between measurement of VE and P2G for forward flexion measurements, no significant correlation ($p > 0.05$).

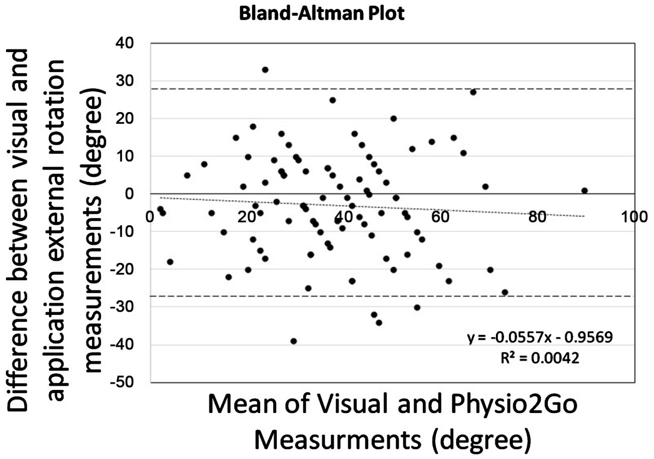

Figure 8. Bland-Altman plot of averages vs. differences between VE and P2G for external rotation measurements, no significant correlation ($p > 0.05$).

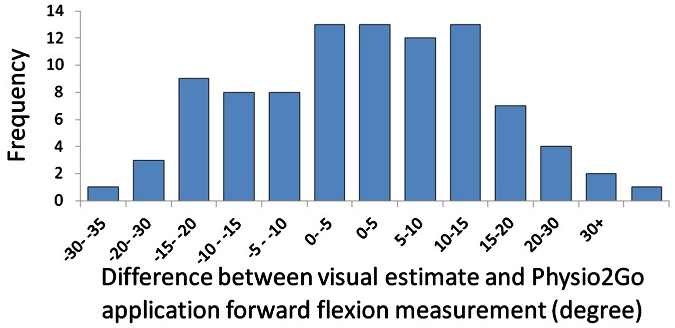

Figure 9. Scatterplot agreement between visual estimation and smartphone application for external rotation measurements ($n = 98$).

In comparison to other studies examining smartphone inclinometer applications, our intra rater reliability measurements for UG (ICC > 0.80) are within the range of other reported values for UG (0.64 - 0.91) [7] [15]. Although no other study has looked at P2G, intra rater reliability measures for smartphone based inclinometers and digital goniometers (ICC > 0.90) [7] [12] [14] are similar to the reliability measures we have obtained (ICC = 0.84 - 0.90). Of note, our ICC intra-rater reliability measures far exceed those recorded using still photography for shoulder ROM (ICC = 0.54) [8]. Additionally, for both research fellow and research assistant we observed

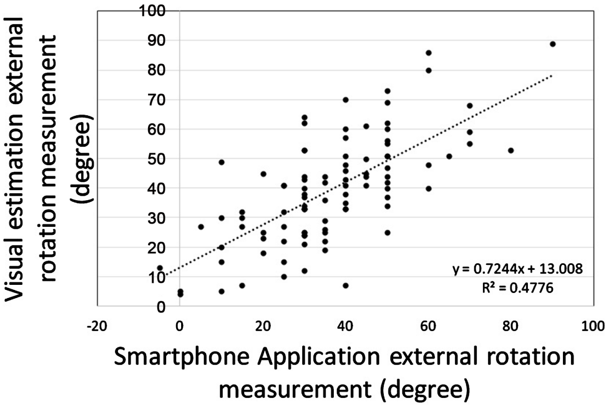

Figure 10. Histogram depiction of difference between visual estimation and Physio2Go external rotation measurement (*n* = 98).

higher intra rater reliability when using the P2G than when using the UG, a trend observed in multiple other studies [7] [14] [15], indicating the ability to obtain repeatable and reliable measures when using a smartphone application. Similarly, our SEM values are well below the clinically acceptable interval of 10˚ [17], with lower SEM achieved when using P2G.

Our reliability values for intra rater across UG and P2G showed substantial or near perfect correlation (ICC > 0.79). Correspondingly, the SEM was low for both observers (<2.5˚), with the lower SEM obtained by the research assistant. Studies using multiple evaluators of varying experience level have also shown similar results that smartphone applications and UG can have high reliability regardless of observer experience level, with the least experienced evaluator occasionally recording the best reliability [15].

The final measure of validity of P2G was the inter-observer reliability measured in forward flexion. Although the UG showed slightly higher inter-observer reliability than P2G (ICC = 0.73), both indicated strongly correlated results. Werner *et al*. reported very similar interrater ICC values for the smartphone inclinometer application they examined (ICC = 0.75) [15], as did Shin *et al*. (ICC = 0.63 - 0.83) [7]. Both of these reliability measures are higher than interrater reliability reported for goniometery (ICC = 0.69) and still photography (ICC = 0.73) [8].

Our SEM for both methods of evaluation was lower (<2.0˚) than those reported for other smartphone ROM applications (7.8˚ - 14.15˚) [7]. In order to confirm our validity we measured PCC for interobserver measurements for both evaluators and methods. The high PCC values (PCC > 0.6) were both significant ($p < 0.05$), indicating strong correlation between UG and P2G ROM measurements for both observers.

High intra rater reliability measures indicated that the raters, irrespective of experience level, were able to obtain reliable and reproducible ROM values for healthy subjects. The research assistant's higher ICC values when using P2G may be due to more time spent familiarizing themselves with the application, although each observer was allowed the same amount of time. Once interrater reliability of P2G was demonstrated the application was then used in the clinical trial comparing P2G and VE in symptomatic shoulder patients. We chose not to validate the ROM measures using ER due to time constraints and the general consensus in literature is that when ICC values for FF are high they are also high for ER measurements [7] [15].

The attending surgeon did not complete a validation for VE as we aimed to demonstrate the applicability of incorporating this application into a busy clinical practice. We expect that as found in previous studies, visual ROM estimates would have low interrater reliability and only moderate intra rater reliability for ROM in symptomatic shoulders [3] [7]. Were the attending to re-familiarize themselves with the visual ROM scale this would negate the clinical relevance; we believe that a visual re-familiarization is not done frequently in practices. A benefit of using a digital application such as P2G is that one doesn't have to re-familiarize themselves with the technique. If used consistently, ROM measurements from a digital application will remain consistent and the scale never becomes faulty.

Due to the necessity to be present at all clinics, only the research assistant performed the P2G measurements for symptomatic patients. The high ICC values obtained by the assistant during validity testing lend credit to their ability to accurately measure using P2G. Normally in the clinical setting a physician assistant or a clinical fellow evaluates a patient first, visually estimating ROM, and the surgeon then visually confirms the recorded ROM. The use of an application to measure ROM by a research assistant is akin to a physician assistant who has the time to learn the technology and measure during their in-office evaluation. IR was not measured because our clinical protocol is to use vertebrae levels and not degrees as indices and thus this measure cannot be performed using a goniometer or virtual goniometer application [19].

As anticipated, there was large disparity between VE and the calculated P2G ROM values. Although ICC values for FF indicate a high degree of correlation (ICC = 0.88), correlation does not necessarily imply agreement. A lack of agreement in FF measurements can be seen as the differences between measurements ranged from 1˚ to 27˚. We also obtained a high SEM (8.9˚), which while still clinically reliable is much higher than the SEM values obtained for interrater reliability during validation.

We believe the apparent "reliability" between measures is due to the large sample size while the low mean difference between measurements is due to the VE values both greater and less than P2G values, balancing out the average. The actual disparity between measures can best be seen in **Figure 5** and **Figure 6**, which show the low agreement between measures, and the high frequency of measures that disagree by more than 10˚. The Bland-Altman plot for FF shows a very wide 95% LOA (3.3˚ ± 23.9˚), indicating that, as expected, the two methods of measurement cannot be considered equivalent (**Figure 7**).

The low agreement between VE and P2G was also apparent for ER. Inter rater comparison had the lowest ICC value measured (0.68), and the correspondingly largest SEM (10.0˚). **Figure 8** and **Figure 10** demonstrate visually the large disparity between the VE and P2G measurements. In the histogram (**Figure 9**), one can see that approximately half (48) of the ER measurements differed by greater than 10˚, thus outside the limit for clinical relevance. The Bland-Altman plot shows the extremely large 95% LOA (−3.1˚ ± 27.4˚), indicating that the two methods for measuring ER are not equivalent (**Figure 8**).

To understand the importance of having both reliable and accurate measures we applied our results to a patient cohort, to mimic what would be done for a research cohort examining patient outcomes. We randomly selected half of our reverse total shoulder patients from the ROM measuring cohort to analyze outcomes. We compared FF values recorded preoperatively (VE by physician assistant) to postoperative values measured by the two methods (VE and P2G). When evaluating change of ROM using the visually estimated postoperative data it appeared that there was a significant improvement in FF (p = 0.007). However, when comparing postoperative outcomes using the P2G measurements there was much less significance (p = 0.02). While this was a small sample size of patients it demonstrated the necessity to use the most reliable method, in this case, P2G, in order to truly validate significance. For ER, neither comparison (pre-to-postoperative) reached statistical significance. Yet still there was again a disparity between significance levels reported for VE and P2G, p = 0.208 and p = 0.898 respectively.

We believe the limits of this study to be due to the time restriction; as we are evaluating a tool that may be updated frequently we wanted to complete this study in a 6 month window to present the most current and up-to-date technology available. Therefore we were unable to use ER for validation, but, as mentioned previously, the literature supports our conclusions made using FF. Additionally this study does not address a method for accurately measuring IR using a smartphone application. We believe further study should investigate IR using smartphone applications, with a focus on quick, simple, methods that don't force the patient into difficult or stressful positions.

This study validates the use of P2G, a novel smartphone application, as an accurate clinical tool employing a virtual goniometer to measure ROM. Having a digital measure of the image would allow the clinician to seamlessly work with other members of the patient care team, such as the physical therapists or physician assistants. P2G required minimal familiarization and allowed clinicians to accurately measure ROM without relying on gravitational calibrations or concerns of body habitus interfering with values, both of which are concerns for inclinometer based applications [7] [15].

As physicians strive to meet the demands of the increasing patient volume, maintain standards of patient care, and improve evaluation measures, they need the best tools available. It is evident from the large amount of clinicians using smartphone technology in their practices that this is an important and emerging aspect of modern medicine. The necessity to accurately measure shoulder ROM to properly monitor postoperative outcomes is critical for both research purposes as well as clinical evaluation. The ease of misreporting these values and the impact this can have was demonstrated through our clinical application showing false significance when VE was used.

The purpose of this study was to investigate the applicability, and efficacy, of a new method of measuring shoulder ROM aimed at improving accuracy for clinical investigators. Our study found that this application provides superior reliability compared to the customary visual physical exam routinely used for orthopedic research. Furthermore our results showed the extent to which research validation can be jeopardized when using physical exam by inspection instead of more accurate methods. The value of using this application instead of a UG is the ease of use and the ability for any member of the healthcare team, regardless of clinical experience to be able to produce reliable, valid measurements, as indicated by the high reliability achieved by the research assistant. We encourage the use of digital ROM applications for any practice that frequently needs quick, recordable values. The use of a standardized application, across orthopedic practices, will help validate research significance, effectively eliminating the existing discrepancies between current measurements. We believe that using this readily available, novel, and innovative technology provides the clinician with the necessary means to effectively and efficiently substantiate clinical values thus elevating the precision of the derived research.

References

[1] Nomden, J.G., Slagers, A.J., Bergman, G.J.D., Winters, J.C., Kropmans, T.J.B. and Dijkstra, P.U. (2009) Interobserver Reliability of Physical Examination of Shoulder Girdle. *Manual Therapy*, **14**, 152-159. http://dx.doi.org/10.1016/j.math.2008.01.005

[2] O'Neill, B.J., O'Briain, D., Hirpara, K.M., Shaughnesy, M., Yeatman, E.A. and Kaar, T.K. (2013) Digital Photography for Assessment of Shoulder Range of Motion: A Novel Clinical and Research Tool. *International Journal of Shoulder Surgery*, **7**, 23-27. http://dx.doi.org/10.4103/0973-6042.109888

[3] Terwee, C.B., de Winter, A.F., Scholten, R.J., Jans, M.P., Devillé, W., van Schaardenburg, D. and Bouter, L.M. (2005) Interobserver Reproducibility of the Visual Estimation of Range of Motion of the Shoulder. *Archives of Physical Medicine and Rehabilitation*, **86**, 1356-1361. http://dx.doi.org/10.1016/j.apmr.2004.12.031

[4] Gajdosik, R.L. and Bohannon, R.W. (1987) Clinical Measurement of Range of Motion Review of Goniometry Emphasizing Reliability and Validity. *Physical Therapy*, **67**, 1867-1872.

[5] Cools, A.M., De Wilde, L., Van Tongel, A., Ceyssens, C., Ryckewaert, R. and Cambier, D.C. (2014) Measuring Shoulder External and Internal Rotation Strength and Range of Motion: Comprehensive Intra-Rater and Inter-Rater Reliability Study of Several Testing Protocols. *Journal of Shoulder and Elbow Surgery*, **23**, 1454-1461. http://dx.doi.org/10.1016/j.jse.2014.01.006

[6] Riddle, D.L., Rothstein, J.M. and Lamb, R.L. (1987) Goniometric Reliability in a Clinical Setting Shoulder Measurement. *Physical Therapy*, **67**, 668-673.

[7] Shin, S.H., Ro, D.H.; Lee, O.-S., Oh, J.H. and Kim, S.H. (2012) Within-Day Reliability of Shoulder Range of Motion Measurement with a Smartphone. *Manual Therapy*, **17**, 298-304. http://dx.doi.org/10.1016/j.math.2012.02.010

[8] Hayes, K., Walton, J.R., Szomor, Z.R. and Murrell, G.A. (2001) Reliability of Five Methods for Assessing Shoulder Range of Motion. *Australian Journal of Physiotherapy*, **47**, 289-294. http://dx.doi.org/10.1016/S0004-9514(14)60274-9

[9] Moreno-Pérez, V., Moreside, J., Barbado, D. and Vera-Garcia, F.J. (2015) Comparison of Shoulder Rotation Range of Motion in Professional Tennis Players with and without History of Shoulder Pain. *Manual Therapy*, **20**, 313-318. http://dx.doi.org/10.1016/j.math.2014.10.008

[10] de Winter, A.F., Heemskerk, M.A.M.B., Terwee, C.B., Jans, M.P.; Devillé, W., van Schaardenburg, *et al*. (2004) Inter-Observer Reproducibility of Measurements of Range of Motion in Patients with Shoulder Pain Using a Digital Inclinometer. *BMC Musculoskeletal Disorders*, **5**, 18. http://dx.doi.org/10.1186/1471-2474-5-18

[11] Epocrates (2013) Epocrates 2013 Mobile Trends Report: Maximizing Multi-Screen Engagement among Clinicians.

[12] Ferriero, G., Sartorio, F., Foti, C., Primavera, D., Brigatti, E. and Vercelli, S. (2011) Reliability of a New Application for Smartphones (DrGoniometer) for Elbow Angle Measurement. *PM&R*, **3**, 1153-1154. http://dx.doi.org/10.1016/j.pmrj.2011.05.014

[13] Johnson, L.B., Sumner, S., Duong, T., Yan, P., Bajcsy, R., Abresch, R., *et al*. (2015) Validity and Reliability of Smartphone Magnetometer-Based Goniometer Evaluation of Shoulder Abduction—A Pilot Study. *Manual Therapy*. [Epub ahead of print] http://dx.doi.org/10.1016/j.math.2015.03.004

[14] Milanese, S., Gordon, S., Buettner, P., Flavell, C., Ruston, S., Coe, D., *et al*. (2014) Reliability and Concurrent Validity of Knee Angle Measurement: Smart Phone App versus Universal Goniometer Used by Experienced and Novice Clinicians. *Manual Therapy*, **19**, 569-574. http://dx.doi.org/10.1016/j.math.2014.05.009

[15] Werner, B.C., Holzgrefe, R.E., Griffin, J.W., Lyons, M.L., Cosgrove, C.T., Hart, J.M. and Brockmeier, S.F. (2014)

Validation of an Innovative Method of Shoulder Range-of-Motion Measurement Using a Smartphone Clinometer Application. *Journal of Shoulder and Elbow Surgery*, **23**, e275-e282. http://dx.doi.org/10.1016/j.jse.2014.02.030

[16] Reese, N.B. and Bandy, W.D. (2013) Joint Range of Motion and Muscle Length Testing. Elsevier Health Sciences, Amsterdam.

[17] Boone, D.C., Azen, S.P., Lin, C.M., Spence, C., Baron, C. and Lee, L. (1978) Reliability of Goniometric Measurements. *Physical Therapy*, **58**, 1355-1360.

[18] Landis, J.R. and Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics*, **33**, 159-174. http://dx.doi.org/10.2307/2529310

[19] Smith, A.M., Barnes, S.A., Sperling, J.W., Farrell, C.M., Cummings, J.D. and Cofield, R.H. (2006) Patient and Physician-Assessed Shoulder Function after Arthroplasty. *The Journal of Bone and Joint Surgery* (*American*), **88**, 508-513. http://dx.doi.org/10.2106/JBJS.E.00132