Assessing Arthroscopic Skills Using Wireless Elbow-Worn Motion Sensors

Georgina S.J. Kirby, BSc(Hons), MRes, Paul Guyver, FRCS(Tr&Orth), Louise Strickland, MSc, Abtin Alvand, DPhil, Guang-Zhong Yang, PhD, FREng, Caroline Hargrove, PhD, Benny P.L. Lo, PhD, and Jonathan L. Rees, MD, FRCS(Tr&Orth)

Investigation performed at the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Science at the Botnar Research Centre and National Institute for Health Research (NIHR) Biomedical Research Unit, University of Oxford, Oxford, United Kingdom

Background: Assessment of surgical skill is a critical component of surgical training. Approaches to assessment remain predominantly subjective, although more objective measures such as Global Rating Scales are in use. This study aimed to validate the use of elbow-worn, wireless, miniaturized motion sensors to assess the technical skill of trainees performing arthroscopic procedures in a simulated environment.

Methods: Thirty participants were divided into three groups on the basis of their surgical experience: novices (n = 15), intermediates (n = 10), and experts (n = 5). All participants performed three standardized tasks on an arthroscopic virtual reality simulator while wearing wireless wrist and elbow motion sensors. Video output was recorded and a validated Global Rating Scale was used to assess performance; dexterity metrics were recorded from the simulator. Finally, live motion data were recorded via Bluetooth from the wireless wrist and elbow motion sensors and custom algorithms produced an arthroscopic performance score.

Results: Construct validity was demonstrated for all tasks, with Global Rating Scale scores and virtual reality output metrics showing significant differences between novices, intermediates, and experts (p < 0.001). The correlation of the virtual reality path length to the number of hand movements calculated from the wireless sensors was very high (p < 0.001). A comparison of the arthroscopic performance score levels with virtual reality output metrics also showed highly significant differences (p < 0.01). Comparisons of the arthroscopic performance score levels with the Global Rating Scale scores showed strong and highly significant correlations (p < 0.001) for both sensor locations, but those of the elbow-worn sensors were stronger and more significant (p < 0.001) than those of the wrist-worn sensors.

Conclusions: A new wireless assessment of surgical performance system for objective assessment of surgical skills has proven valid for assessing arthroscopic skills. The elbow-worn sensors were shown to achieve an accurate assessment of surgical dexterity and performance.

continued
Dependent arthroscopies were classiﬁed as novices (n = 15), those who had performed fewer than 100 independent arthroscopies were classiﬁed as intermediates (n = 10), and those who had performed 100 or more independent arthroscopies were classiﬁed as experts (n = 5). For each participant, the total number of arthroscopic procedures previously performed was recorded from the participant’s surgical logbooks. None of the participants had any previous experience of the simulator or the virtual reality tasks.

**Clinical Relevance:** The validation of an entirely objective assessment of arthroscopic skill with wireless elbow-worn motion sensors introduces, for the first time, a feasible assessment system for the live operating theater with the added potential to be applied to other surgical and interventional specialties.

Although surgical competence consists of a number of domains such as decision making, technical dexterity, communication, and knowledge, technical dexterity is still considered one of the essential attributes of a surgeon and underperformance in this domain can lead to intraoperative errors, excess operative time, and harm to patients. The acquisition of technical skills therefore remains a fundamental goal of surgical training. However, the restructuring of modern surgical education from the traditional time-based apprenticeship, together with restrictions on trainee’s working hours in Europe and North America, has resulted in a dramatic reduction in the time available to train a surgeon and the number of operative cases that they may perform.

Consequently, new approaches to teaching and assessing surgical skills with training that takes place outside the operating theater are evolving. Surgical skills training with use of simulators is becoming more available, allowing trainees to learn and to gain experience without compromising patient safety. Simulators can be employed for teaching a number of surgical skills in a safe, standardized, and adaptable environment. There is also opportunity for assessment and feedback, with new methods of assessing and quantifying surgical skill having been developed.

Motion analysis systems have been used in simulation centers to objectively measure technical dexterity and learning of surgical skills, but it has not been possible to use them in the operating room because most are tethered by wires and are based on electromagnetic tracking. As such, assessment in the real operating theater is limited to paper methods such as Global Rating Scales and checklists. These are time-consuming but, at present, are the only methods practical for use in the operating theater.

We recently developed a new wireless method of objectively assessing surgical dexterity. To date, this method has only been tested with wrist sensors, limiting its use. The aim of this study was to test and to validate this novel wireless motion system for use with small sensors worn on the elbow. Validation of this method would allow, for the first time, truly objective assessment of trainee surgical skills in the real operating room. In turn, this would facilitate the development of evidence-based simulation training.

**Materials and Methods**

**Settings and Subjects**

This work was carried out in a dedicated orthopaedic simulation center in a university teaching hospital with ethical approval for this educational study. Participation was entirely voluntary and thirty participants were enrolled and were split into three groups (novice, intermediate, and expert) depending on their level of arthroscopic experience. Those who had not performed any independent arthroscopies were classiﬁed as novices (n = 15), those who had performed fewer than 100 independent arthroscopies were classiﬁed as intermediates (n = 10), and those who had performed 100 or more independent arthroscopies were classiﬁed as experts (n = 5). For each participant, the total number of arthroscopic procedures previously performed was recorded from the participant’s surgical logbooks. None of the participants had any previous experience of the simulator or the virtual reality tasks.

**Shoulder Arthroscopy Simulator Tasks**

A VirtaMed ArthroS virtual reality simulator (VirtaMed, Zurich, Switzerland) with a 30° arthroscope was used for this study, with all participants performing three tasks: Diagnostic I (a standardized visual examination of ten anatomical landmarks in the shoulder joint), Triangulation I (a standardized probe examination of five spheres in the shoulder joint), and Triangulation II (the hooking of ﬁve rings in the shoulder joint). Figure 1 shows the setup of the simulator. Prior to undertaking the tasks, all participants received standardized instructions and a presentation introducing the simulator and the tasks to be performed. They were then given a few minutes to familiarize themselves with the equipment.

**Motion Sensor Setup**

Motion analysis was performed with use of a newly developed wireless assessment of surgical performance (WASP) system, a combination of motion sensors built with a 30° arthroscope used for this study, with all participants performing three tasks: Diagnostic I (a standardized visual examination of ten anatomical landmarks in the shoulder joint), Triangulation I (a standardized probe examination of five spheres in the shoulder joint), and Triangulation II (the hooking of five rings in the shoulder joint). Figure 1 shows the setup of the simulator. Prior to undertaking the tasks, all participants received standardized instructions and a presentation introducing the simulator and the tasks to be performed. They were then given a few minutes to familiarize themselves with the equipment.

**Fig. 1** VirtaMed ArthroS virtual reality simulator setup with motion sensors on wrists and elbows of the subject.
TABLE I Task Performance Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Right-hand dominant (%)</th>
<th>Right-hand stationary (%)</th>
<th>Left-hand dominant (%)</th>
<th>Left-hand stationary (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time taken</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoothness of movement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of hand movements</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall maximum hand acceleration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range of accelerometer values</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Range of gyroscope values*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance of resultant acceleration*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance of resultant gyroscope values*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity of movement in the 0 to 5-Hz frequency band*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity of movement in the 5 to 20-Hz frequency band*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*These parameters are calculated for both the left and right hands.

and custom algorithms for deriving performance parameters and an overall arthroscopic performance score. Four commercially available miniaturized motion sensors (WAX9; Axivity, York, United Kingdom) were set up to collect data from their three-axis accelerometer and three-axis gyroscope at a frequency of 50 Hz; they also contain a three-axis magnetometer that was not used because of the magnetic field distortions that would occur in the majority of clinical settings. These sensors have up to eight hours of continuous battery life, are USB (universal serial bus)-rechargeable, and can stream their outputs over a Bluetooth connection. There were therefore six data streams from each of the four devices being simultaneously streamed live over Bluetooth and being recorded on a dedicated laptop. Because the sensors do not have an output of time, they were time-synchronized by detecting similar features in the accelerometer data. Customized software was written in MATLAB 2014 (MathWorks, Natick, Massachusetts). MATLAB was also used for all data processing and statistical analysis.

Determining Sensor Locations

For this study, sensors were worn simultaneously on both wrists and both elbows. The wrist sensor location was already validated in previous work on simulated laparoscopy and is the midpoint of the dorsum of each wrist, level (or in line) with the ulnar styloid. To use such sensors in the real operating theater, they would need to be worn close to the elbow to allow the surgeon to scrub. An initial pilot experiment was therefore conducted and a standardized and reproducible position to wear the elbow sensors was three fingerbreadths below the lateral epicondyle in line with the radius. This position was sufficiently proximal to allow for future use in the operating room, but distal enough from the elbow to be influenced by contractile movements of the extensor muscle bellies during arthroscopic surgical activity.

Data Output and Outcome Measures

The following data outputs were recorded for each participant for each task: ArthroS virtual reality simulator task video recordings and objective output metrics (time taken and camera and instrument path lengths in centimeters) and objective motion and rotation data collected and recorded over Bluetooth from four wireless motion sensors (one on each wrist and one on each elbow).

Global Rating Scale

A previously validated Global Rating Scale (GRS) was used to assess each individual task video. This was undertaken by an assessor blinded to participant identity. Overall scores were used as a gold standard for comparison with the wireless motion data. Although the GRS is well validated, we ensured ongoing interobserver reliability with another blinded assessor rescoring eighteen videos. The GRS again had excellent interobserver reliability, with a Cronbach alpha of 0.95.

Arthroscopic Performance Score Using Wireless Motion Analysis

Customized scripts in MATLAB converted the recorded wireless motion data into a number of performance parameters (Table I). A new parameter, the smoothness of movement, was also derived from the accelerometer signal by calculating the change in overall acceleration for each hand.

The performance parameters derived from the data for each participant were input into an unsupervised k-means clustering technique; this technique has previously been used for the assessment of skill on a laparoscopic task. For each of the tasks, the output of this machine learning technique gave three distinct groups of participants, clustering together those who have similar performance. An arthroscopic performance score (APS) was assigned to each group on the basis of performance on the task: APS 1 was assigned to the lowest performing group, APS 2 was assigned to the intermediate performing group, and APS 3 was assigned to the highest performing group.

Two arthroscopic performance scores were determined for each participant from the motion data, one from the wrist-worn sensors and one from the elbow-worn sensors. These arthroscopic performance scores were then compared with both the GRS gold-standard scores and the ArthroS simulator metrics with the purpose of validating the use of the elbow wireless motion sensors and their customized outputs as a successful and objective measure of surgical performance on these arthroscopic tasks.

TABLE II Correlations Between GRS Gold-Standard Scores and Motion-Derived Arthroscopic Performance Scores for Each Arthroscopic Task

<table>
<thead>
<tr>
<th>Task</th>
<th>Wrist Arthroscopic Performance Score and GRS Score</th>
<th>Elbow Arthroscopic Performance Score and GRS Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic I</td>
<td>0.62</td>
<td>0.76</td>
</tr>
<tr>
<td>Triangulation I</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Triangulation II</td>
<td>0.73</td>
<td>0.80</td>
</tr>
</tbody>
</table>

*All of these corrections were significant at p < 0.001.
Statistical Analysis
Although the GRS scores were a continuous variable, the Shapiro-Wilk test demonstrated that these scores were not normally distributed; therefore, non-parametric tests were applied.

Comparison of GRS with Arthroscopic Performance Score
Spearman rank correlation analysis was conducted between the GRS score of each participant for each task and the two novel motion-derived performance scores (one arthroscopic performance score for each sensor location).

Comparison of Virtual Reality Output Metrics with Arthroscopic Performance Score
A Kruskal-Wallis test was used to test for significant differences in the total path-length output from the ArthroS simulator (instrument and camera path lengths combined) for each motion-derived arthroscopic performance score level. The correlation between the total path-length output from the virtual reality simulator and the number of hand movements derived from the motion sensors was determined with use of the Pearson correlation coefficient.

Comparison of the Performance Parameters
The performance parameters derived from both the wrist-worn and elbow-worn sensor locations (as listed in Table I) were tested for significant differences between the distinct arthroscopic performance score levels (APS 1, APS 2, and APS 3) with use of a Kruskal-Wallis test. For all statistical analysis, significance was set at $p < 0.01$.

Source of Funding
One author of this study (G.S.J.K.) received an education grant from McLaren Applied Technologies, which is also her employer; these funds were used for the university fees associated with her thesis, part of which involved this study. The Oxford NIHR Biomedical Research Unit provided infrastructure support for this study. The ArthroS virtual reality simulator was loaned for one year to the University of Oxford by VirtaMed, Zurich, Switzerland.

Results
The ArthroS virtual reality output metrics demonstrated construct validity for each of the three arthroscopic tasks (Fig. 2), with the novices, intermediates, and experts showing the expected significant differences in time taken and total path length ($p < 0.001$ for all; Kruskal-Wallis test). The GRS scores also confirmed construct validity for each of the three arthroscopic tasks, with significant differences again between the novice, intermediate, and expert groups (Kruskal-Wallis test, with all $p < 0.001$).

Wireless Wrist-Worn Sensor Arthroscopic Performance Score
For each distinct cluster output from the k-means clustering technique, an arthroscopic performance score (APS 1, APS 2, or APS 3) was assigned by determining the level of expertise in that group; the members of the lowest-performing cluster on that particular task were determined to have APS 1 and the cluster of highest performers were determined to have APS 3.

Comparison of Wrist Sensor Arthroscopic Performance Scores with GRS Scores
The wrist sensor arthroscopic performance scores and GRS scores showed significantly high correlations ($p < 0.001$) for the Triangulation I and II tasks with a slightly lower, but still significant, correlation ($p < 0.001$) for the Diagnostic I task (Table II). The number of hand movements derived from the wireless

Fig. 2
Boxplot showing the significant differences in total path length for the Triangulation II task between participants of differing surgical expertise. The crosses denote outliers and the asterisks denote significant differences ($p < 0.001$). The box represents the interquartile range, the red line in the box represents the median, and the whiskers represent the 2.5 and 97.5 percentiles.
wrist-worn motion sensors also differed significantly \((p < 0.01)\) between the different levels of expertise on the arthroscopic performance score (APS 1, APS 2, and APS 3) for all three tasks; see the left plot in Figure 3.

**Comparison of Wrist Sensor Arthroscopic Performance Scores with Virtual Reality Output Metrics**

As with the above validation of the arthroscopic performance score to the GRS, similar significant findings \((p < 0.01)\) were present when the ArthroS simulator path-length output was compared with the wrist motion sensor performance levels (APS 1, APS 2, and APS 3) (Fig. 4). The total path-length output from the virtual reality simulator and the number of hand movements derived from the wrist sensors for each task had significantly high correlations \((p < 0.001)\) on the Diagnostic I task (0.91), the Triangulation I task (0.94), and the Triangulation II task (0.92).

**Wireless Elbow-Worn Arthroscopic Performance Score Validity**

For the elbow sensor motion output, the arthroscopic performance scores were again assigned to the three distinct clusters and comparisons with GRS gold-standard scores and ArthroS virtual reality output metrics were similarly examined.

**Comparison of Elbow Arthroscopic Performance Scores with GRS Gold-Standard Scores**

The elbow sensor arthroscopic performance scores and GRS scores showed significantly high correlations \((p < 0.01)\) for all three arthroscopic tasks (Table II).

The number of hand movements also showed significant differences \((p < 0.01)\) between the three levels of expertise (APS 1, APS 2, and APS 3) over all three arthroscopic tasks; see the right plot in Figure 3.

**Comparison of Elbow Arthroscopic Performance Scores with Virtual Reality Output Metrics**

The total path-length output from the ArthroS simulator (instrument and camera path lengths combined) showed significant differences between APS 1, APS 2, and APS 3 for the Diagnostic I, Triangulation I, and Triangulation II tasks \((p < 0.01)\).

---

**TABLE III Correlations Among Motion Sensor Parameters Derived from Wrist-Worn Sensors and Elbow-Worn Sensors**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Correlation*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (s)</td>
<td>1</td>
</tr>
<tr>
<td>No. of hand movements</td>
<td>0.998</td>
</tr>
<tr>
<td>Smoothness of movement</td>
<td>0.91</td>
</tr>
<tr>
<td>Maximum hand acceleration in g</td>
<td>0.74</td>
</tr>
</tbody>
</table>

*All of these correlations were significant at \(p < 0.001\).
Fig. 4
Boxplots showing significant differences in the Arthros total path-length outputs (in centimeters) from the virtual reality simulator taken to complete the Triangulation II task between APS 1, APS 2, and APS 3 for each sensor location. The crosses denote outliers and the asterisks denote significant differences ($p < 0.01$). The box represents the interquartile range, the red line in the box represents the median, and the whiskers represent the 2.5 and 97.5 percentiles.

Fig. 5
Boxplots showing how the number of hand movements changes between the wrist-worn and elbow-worn sensors for each of the three tasks; task 1 refers to the Diagnostic I task, task 2 refers to the Triangulation I task, and task 3 refers to the Triangulation II task. The crosses denote outliers and the asterisks denote significant differences ($p < 0.01$). The box represents the interquartile range, the red line in the box represents the median, and the whiskers represent the 2.5 and 97.5 percentiles.
The total path-length output from the virtual reality simulator and the number of hand movements derived from the elbow sensors for each task were also found to be significantly correlated (p < 0.01) for the Diagnostic I task (0.91), the Triangulation I task (0.93), and the Triangulation II task (0.93). The plots in Figure 4 demonstrate how the path length that was moved during the Triangulation II task changed between participants clustered in APS 1, APS 2, and APS 3.

Comparison of Wrist-Worn and Elbow-Worn Sensor Arthroscopic Performance Scores

The motion sensor parameters showed significantly high correlations between the wrist-worn and elbow-worn locations (p < 0.001), demonstrating that the elbow-worn sensors measure metrics similar to those derived from the validated wrist sensors (Table III). The number of hand movements derived from the wireless motion sensors also differed significantly (p < 0.01) between the three arthroscopic tasks (Fig. 5).

The GRS score for each participant was then compared with the arthroscopic performance score derived for both the wrist-worn and the elbow-worn sensor outputs (Tables II and III). The elbow-worn sensor arthroscopic performance score had significantly higher correlations (p < 0.001) to the GRS score than did the wrist-worn sensor arthroscopic performance score.

Further detailed analysis of individual arthroscopic performance score classifications was also undertaken. Figure 6 shows the cluster assignments for the Triangulation II task in which each participant is plotted against just three of the input features: time taken, smoothness of movement, and number of hand movements. In Figure 6, the left graph demonstrates the arthroscopic performance score output using the wrist-worn sensor data and the right graph shows the arthroscopic performance score output using the elbow-worn sensor data. Highlighted in green data points in Figure 6 is the cluster of experts. It can be seen that a number of APS 2 data points had been wrongly clustered to APS 3 when using features derived from the wrist sensor data, but they had been correctly assigned to APS 2 when data from the elbow sensors were used. This means that wearing sensors on the elbow more accurately detects a participant’s level of expertise on the task than wearing them on the wrist.

The same results were obtained for the other two tasks with the elbow sensors; the APS 3 group contained the highest performers who demonstrated significantly shorter task time, significantly fewer hand movements to complete the task, and significantly lower maximum accelerations (p < 0.001 for all).

Discussion

With the growing importance of teaching, learning, and assessing surgical skills, this study has successfully demonstrated the validity of a new objective assessment system that is entirely wireless. Although one aim was to develop a wireless system that could be used with wrist-worn sensors (that would be useful in a simulation setting), the main aim was to assess the validity of a wireless system that utilized elbow-worn sensors. The main output from this new wireless method is the arthroscopic performance score. The clustering technique used in calculating the arthroscopic performance score has already been used in previous work on simulated laparoscopy and has again proven valid and useful as a categorization technique for converting wireless objective motion parameters into performance level scores. What will surprise many is that the elbow-worn sensors not only proved a feasible and valid method for assessing dexterity skills, but also actually showed greater correlation to surgical performance than the wrist-worn sensors. The likely reason for this is the
consistent placement of the elbow sensors over the common extensor muscle bellies in the proximal aspect of the forearm. Besides movement detection with positional change of the limb in a three-dimensional space, it would also result in motion detection with contractions of the extensor muscle bellies during movements of the surgeon’s hands and wrists. These results at present are only relevant to arthroscopic surgery, and further data collection, analysis, and validation would be needed for open surgery.

A limitation of this study might be the number of study participants. A larger number of data points from intermediate participants with more varied expertise would enable further cluster analysis. With more data points, there may be the possibility of even more distinct clusters and the prospect of more arthroscopic performance score levels providing further granularity to the performance scores.

If surgical skill assessment can move away from requiring expert surgeons to complete time-consuming checklists and rating scales on trainees and can move toward cost-effective, reliable, and user-friendly objective assessment methods, then the advantages are clear. Furthermore, such systems offer users the option to practice regularly in safe simulated settings, while receiving real-time objective feedback on performance without the need for an expert assessor. To date, trainees have only been able to receive real-time objective feedback when using virtual reality trainers, which have obvious cost and access limitations. However, motion sensors are considerably cheaper and portable. Thus, using this type of motion sensor system on bench models could provide a more accessible system to trainees and surgeons.

Interestingly, this study also revealed significantly high correlations between the total path-length output from the ArthroS simulator and the number of hand movements derived from the elbow sensor, proving that deriving the number of hand movements directly from the inertial data is sufficient to replace path length for determining task performance.

The most exciting outcome of this study is that surgeons can wear elbow-worn sensors and can scrub in the theater without any adaptations or modifications to the sterile technique (or the sensor can simply be positioned with an elbow strap after scrubbing). This development is important and means, for the first time, that the potential exists to objectively assess surgical skill and dexterity in the real operating room. Following this study, the elbow-worn sensors have already been successfully piloted in the real operating theater for arthroscopic shoulder cases, successfully streaming and collecting data.

We now plan to further validate our system (WASP) to monitor individual simulated learning curves and to conduct an evidence-based assessment of simulation training by utilizing WASP for transfer validation studies to the real operating room.

References

5. E-mail address for J.L. Rees: jonathan.rees@ndorms.ox.ac.uk


