Smoothness of Surgical Tool Tip Motion Correlates to Skill in Endovascular Tasks

Sean Estrada, Cassidy Duran, Daryl Schulz, Jean Bismuth, Michael D. Byrne, and Marcia K. O'Malley, *Senior Member, IEEE*

***Abstract*—Current performance assessment techniques in endovascular surgery are subjective or limited to grading scales based solely on an expert's observation of a novice's task execution. Since most endovascular procedures involve performing fine motor control tasks that require complex dexterous movements, this paper evaluates objective and quantitative metrics of performance that capture movement quality through the computation of tool tip movement smoothness. An experiment was designed that involved recording the catheter tip movement from 20 subjects performing four fundamental endovascular tasks in each of three sessions using manual catheterization on a physical model and in a simulation environment. Several motion-based performance measures that have been shown to reliably assess skill in other domains were computed and tested for correlation with subjective data that were simultaneously obtained from the global rating scale assessment tool. Metrics that captured movement smoothness produced statistically significant correlations with the observation-based assessment metrics and were able to differentiate skill among participants. In particular, submovement analysis led to metrics that captured statistically significant differences across ability group, session, experimental platform, and task. Objective and quantitative metrics that capture movement smoothness could be incorporated into future training protocols to provide detailed feedback on trainee performance.**

***Index Terms*—Motion capture, skill assessment, surgical training, virtual reality.**

I. Introduction

THERE is evidence that skill level can affect clinical outcomes after surgery [1], and the development of effective metrics to evaluate surgical skill is an active area of research [2]. Specifically, the need for *objective* and quantitative assessment tools has been a topic of considerable interest and importance [3]–[5]. To date, structured human grading of surgical skill, wherein rated checklists are employed so that senior surgeons and instructors can evaluate surgical skill of the student, remains the most developed approach for skill evaluation. Criticisms of the approach remain, especially objectivity issues with human observers [2] and the need for experts to conduct the evaluations.

A number of researchers have explored the potential to evaluate surgical tool motions as a means to infer skill level in open and minimally invasive surgery (MIS) (see [2] for a review). The rapid growth of this line of research is due to the expanding availability and use of bench top models and simulators that facilitate the use of or even incorporate motion capture technologies for recording data during practice. For example, the Imperial College Surgical Assessment Device is an electromagnetic (EM) tracking device with custom analysis software that reports number of movements, path length, and completion time to assess performance during laparoscopic surgical procedures [6]. Another system, the Red Dragon, is a mechanical motion tracking system and virtual reality interface that reports kinematics of tools and dynamics of tool/hand interaction for surgical skills assessment [7].

Results from previous studies have demonstrated an ability to assess the skill level of surgeons using quantitative measures derived from motion analysis during surgical procedures on inanimate models and virtual reality simulators. For example, Cotin *et al.* proposed a practical scientific approach to define a set of objective metrics during MIS by tracking the motion of two laparoscopic instruments from a five degree of freedom training device [8]. O'Toole *et al.* observed differences in suturing skill between medical students and surgeons in a virtual reality simulator using tool position and force data [9].

To date, these approaches have not been applied to the endovascular surgical domain. Endovascular surgery is a form of MIS that enables access to many regions of the body via major and minor blood vessels [see Fig. 1(a)]. The surgeon introduces a catheter into a port that gives access to the vasculature, usually via the femoral artery, to perform an array of endovascular diagnostic and therapeutic procedures. To help deliver the catheter into the target blood vessel, a sheath moving over the catheter and guidewire moving through the catheter act as a guide and support [see Fig. 1(b)]. To visualize the movements of these instruments, endovascular surgeons use live X-ray (fluoroscopic) images. The imaging system can be repositioned to provide 2-D projection views of the areas being treated.

Significant differences exist between endovascular surgical techniques and open or laparoscopic surgery that potentially complicate the identification of tool-motion-based performance metrics that correlate to skill. The most notable difference is that the tools used in endovascular surgery are flexible, and the interactions between guidewire, catheter, sheath, and vessel walls significantly affect the motion of the tool tips. Indeed, it has

Manuscript received October 9, 2014; revised November 16, 2015 and February 15, 2016; accepted March 5, 2016. Date of publication April 28, 2016; date of current version September 14, 2016. This paper was recommended by Associate Editor J. Marquard.

S. Estrada and M. K. O'Malley are with the Department of Mechanical Engineering, Rice University, Houston, TX 77005 USA (e-mail: sje1@rice.edu; omalleym@rice.edu).

C. Duran, D. Schulz, and J. Bismuth are with the DeBakey Heart and Vascular Center, Houston Methodist Hospital, Houston, TX 77030 USA (e-mail: caduran@houstonmethodist.org; dgschulz@houstonmethodist.org; jbismuth@houstonmethodist.org).

M. D. Byrne is with the Department of Psychology, Rice University, Houston, TX 77005 USA (e-mail: byrne@rice.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2016.2545247

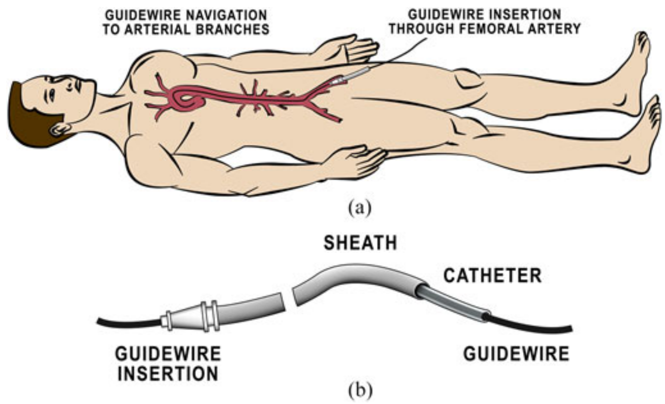


Fig. 1. Endovascular procedure basics. (a) Insert guidewire into femoral artery, navigate to branch of interest. (b) Navigate and steer by varying relative position of guidewire, catheter, and sheath tips, and rotating the precurved guidewire at insertion.

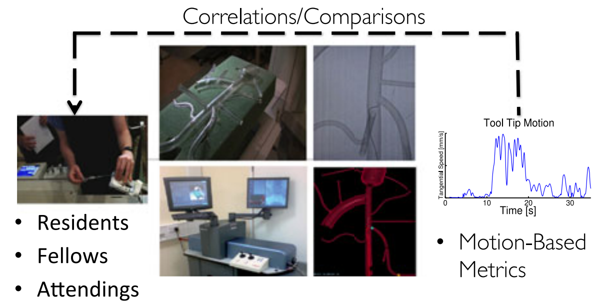


Fig. 2. Approach for identifying motion-based performance metrics of endovascular skills. (Left) Subject performing endovascular task. (Center) Experimental platforms (and fluoroscopic images). (Right) Representative speed trajectory of catheter tip during task execution. Correlations are used to relate the measured motion data to observation-based assessments; then, ability group differences are determined.

been noted that skills should be assessed separately for each surgical paradigm given their unique characteristics [2]. We hypothesize that despite these unique circumstances, performance metrics derived from tool tip motions in endovascular procedures will correlate with surgical skill as has been observed in other domains. Indeed, our pilot investigations showed that motion-based metrics that quantified smoothness were more strongly correlated to the structured grading assessment than those metrics that simply quantified the catheter tool tip kinematics [10].

We take our inspiration for identifying performance metrics from the field of human motor control. Specifically, we assert that motor control inspired measures, such as smoothness of movement, which are derived from the kinematic data of tool tip motion, will give insight into the quality of tool movement [11] and correlate to surgical skill. Movement smoothness is widely regarded as a hallmark of skilled coordinated movement [12], and metrics that capture movement smoothness have been used to assess motor performance in basic motor control tasks [13], [14] and rehabilitation applications [15]–[17]. Others have begun to explore the applicability of human motor control inspired metrics for the surgical domain. Nisky *et al.* showed that the effect of surgical robot manipulators on movements of novices and surgeons could be characterized by a number of metrics including one that quantified the shape of the velocity trajectory (a way to represent smoothness) [18]. Their study was focused on simple pointing movements with a surgical robot, examining the effect of robot dynamics on movements. Their work does not address our challenge of identifying metrics for endovascular tasks that require use of flexible tools. Despite this added complexity for our domain, there is evidence from more fundamental motor control studies that when motor tasks involve flexible manipulated objects, humans adapt their hand motions from the typically observed minimum-jerk trajectory in order to elicit optimally smooth movements of the object [19].

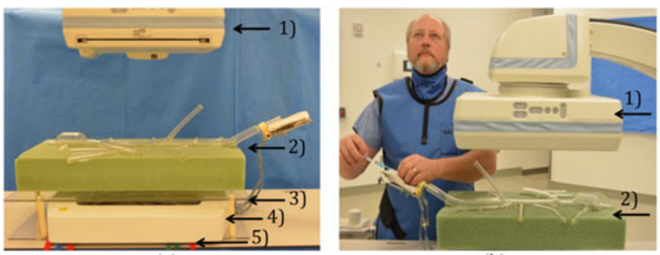
This paper explores the applicability of motion-based measures of performance, such as those previously demonstrated for motor control and rehabilitation, to endovascular surgery procedures. While motion capture techniques have been used in the laparoscopic domain to quantify performance and compare surgical skill, we restrict ourselves to evaluate metrics that capture underlying and fundamental principles of human motor control, specifically aspects of movement smoothness, in order to gain insight into the key characteristics of expert movements. We evaluate performance in surgically relevant tasks specific to the endovascular domain in order to capture the unique characteristics of this specialty. We isolate specific tasks so as to reduce the risk of confounding our findings with assessment of procedural knowledge, and we evaluate performance among domain experts. As illustrated in Fig. 2, catheter tip motion data are collected during the performance of fundamental endovascular tasks on two different platforms (manual catheterization of a physical inanimate model and the same model on a simulator). Simultaneously, assessments using standardized checklists are collected. Correlations and comparison tests are computed in order to identify appropriate motion-based measures that can be used to assess skill for endovascular surgeons. We identify metrics that correlate with structured human grading assessments, that are specific enough to show significant performance differences among domain experts, and that are applicable to multiple training platforms. Further, the ability to compute such performance metrics from data gathered during surgical training offers the possibility to provide specific and directed feedback to the student on the movement strategies they are employing to complete the task. Such metrics are superior to outcome-based performance measures that are not able to be provided as feedback to the student until the task is completed [14].

II. Methods

The experiment was designed to accurately record the movement of the catheter tip while surgeons of different experience levels performed basic endovascular surgical tasks using manual catheterization in a fundamentals of endovascular skills (FEVS) inanimate model and in a simulation environment with a virtual representation of the FEVS model. Subjects were given a predetermined amount of time to complete a series of tasks on both platforms. Table I shows the between-subject and within-subject factors and their associated levels that made up the experimental design.

TABLE I
ALL BETWEEN- AND WITHIN-SUBJECT FACTORS WITH ASSOCIATED LEVELS

Factor	Type	Levels
Platform	Within	Manual and Simulator
Task	Within	Anterior, Posterior, First left lateral, Second left lateral
Session	Within	1, 2, 3
Ability Group	Between	Novice, Intermediate, Expert



(a) (b)

Fig. 3. Experimental setups of catheterization with the physical model. (Left) General setup, with 1) the face of the Siemen's C-arm to generate fluoroscopic images, 2) inanimate physical model with styrofoam housing, 3) plexi-glass casing, 4) window field generator, and, 5) velcro straps to secure setup to patient table. (Right) User performing catheterization using the physical model. (a) Manual catheterization setup. (b) User cannulating model.

In this study, endovascular domain specialists with varying levels of experience completed surgical tasks on a physical inanimate model (see Fig. 3) and in a virtual reality simulator (see Fig. 4). The catheter tip movements, which were manually controlled by the operator, were recorded, and performance metrics were derived from the collected data.

A. Subjects

Twenty subjects (18 male, two female, average age of experts 49, intermediates 39, and novices 32) participated in this study performed at the Houston Methodist Hospital. Seven subjects were either cardiology or vascular residents, six were cardiovascular fellows, six were attending physicians, and one was a lab technician with significant experience performing endovascular procedures in nonclinical scenarios. The subjects ranged in experience from residents less than a year removed from medical school to attending surgeons with more than 20 years of experience in cardiovascular surgery. Novices (ten: six residents, three fellows, one attending) were those having performed less than 50 cases. Intermediates (six: one resident, three fellows, one attending, one lab technician) had performed between 50 and 500 cases. Experts (four: all attending) were defined as participants with more than 500 previous cases. This pool of subjects represents all of the vascular attending surgeons at Houston Methodist, almost all of the vascular residents and all of the vascular fellows training at Methodist during the period of the study, and a small number of attendings, residents, and fellows in closely related specialties who expressed interest in the study.

(a) (b)

Fig. 4. Experimental setups of simulation. (Left) Closeup of Angio Mentor simulator. (Right) User performing catheterization using Angio Mentor simulator. (a) Angio Mentor simulator. (b) User cannulating simulator.

B. Experimental Platforms

The endovascular surgery specialty relies heavily on the use of simulation for training, which can include physical models, part-task trainers, and virtual reality platforms, enabling preprocedural planning, skills acquisition, device training, and maintenance of skills [20]. In this study, we used an anatomically inspired model for manual catheterization. In one case, subjects manually inserted tools into the physical model while visualizing the procedure using fluoroscopy (see Fig. 3). In the other case, subjects manually inserted tools into a virtual reality simulator loaded with a computerized version of the same model (see Fig. 4), visualizing the procedure using simulated fluoroscopy.

In the laparoscopic domain, the fundamentals of laparoscopic surgery program is among the most robust programs available for assessing skill level. The program contains both cognitive and skills components, has been correlated with clinical performance, and offers a certification process through validated examinations [21]. Endovascular surgeons are interested in having a comparable program for the training of skills. Therefore, our group developed the FEVS model to enable task-specific training for endovascular surgery. The FEVS model is nonanatomical (although anatomically inspired) and was designed so that fundamental skills of endovascular surgery could be assessed, including basic catheter and guidewire skills, selective catheterization of all vascular beds, and arteriography of all vascular beds [22], as opposed to specific procedural knowledge and skills.

Virtual reality simulation technology enables the replication of the FEVS inanimate model in computer simulation so that skill can be assessed in both environments, given their widespread use in the specialty. We used the Simbionix ANGIO Mentor Ultimate, which allows users to execute our selected tasks, requiring the insertion and manipulation of guidewires, catheters, and sheaths, into the virtual FEVS model. Guidewire and catheter positions and movements are simulated and are a function both of the surgeons' insertion techniques and rendered anatomical structure. Users control a simulated portable image system and patient table to achieve various fluoroscopic projections. The FEVS model and fluoroscopic image from the surgical simulator from the first left lateral task (described below) are shown in Fig. 5.

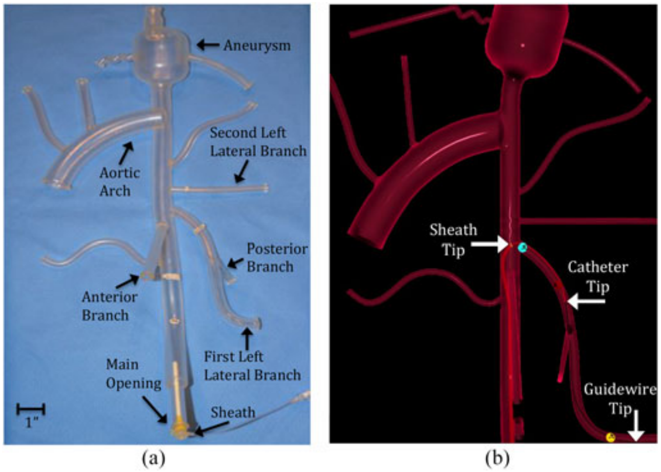


Fig. 5. (a) Inanimate FEVS model. (b) Simulator visualization of the inanimate model with sheath, catheter, and guidewire tips in view. (a) Physical model. (b) Model in simulator.

C. Tasks

Subjects performed four fundamental endovascular tasks on each platform repeated over three sessions. The tasks represent different target locations in the branches of the anatomy, specifically catheterization into the anterior, first left lateral, posterior, and second left lateral branches in the physical or simulated FEVS model. The tip of the guidewire is curved, such that rotation of the guidewire (by rolling it in the fingertips at the entrance to the port) and control of the relative position of the catheter, sheath, and guidewire tips (by advancing them independently at the port entrance) dictate the trajectory of the tools. The order of platform and tasks performed on each platform were pseudorandomized to mitigate learning effects. The first and second left lateral branches were cannulated, while the portable imaging system was at 0°, or, anterior/posterior (A/P), the anterior branch was cannulated at 75° left anterior oblique (LAO), and the posterior branch was cannulated at A/P until the catheter was sufficiently inside of the first left lateral branch, where the rest of the task was performed at 75° LAO.

The goal was to either 1) successfully cannulate the catheter tip to an identified success point in the physical FEVS model (between 2 and 4 cm inside of the branch of interest) within 5 min or 2) successfully cannulate the tips of the guidewire, catheter, and sheath to color-coded targets inside of the branch of interest within five minutes using the simulator. The slight discrepancies between these objectives are due to limitations of our tool tracking technology in the physical model (we can track the tip of only the catheter) versus the commercial simulator which required the three tools to reach their respective targets. While accomplishing all tasks, the guidewire was required to be cannulated into the branch of interest prior to the catheter so that the catheter would go into a branch over the guidewire. If the subject failed to do so, the error was deemed critical because failure to move a catheter over a guidewire in actual endovascular procedures could cause severe complications. When such failures occurred, the corresponding assessment data were dealt with as outlined in Section III-A.

D. Motion Analysis

The kinematic movement of the catheter tip was recorded using an EM tracking device during manual intervention on the physical model and using video-processing techniques for tasks performed on the endovascular simulator. A Northern Digital Incorporated Window Field Generator was used to record the position and orientation of the catheter tip during manual intervention on the physical model due to its ability to track flexible surgical tools while minimizing interference from live fluoroscopy to the generated electric field. The x, y, and z position and orientation in the x- and y-directions were read from the movement of the catheter tip during each task performed on the physical model at a rate of 40 Hz by integrating a single five-degree-of-freedom sensor into each of the catheters. EM tracking is widely accepted and minimally impacts clinical procedures such as ours [23], [24]. When subjects performed procedures on the ANGIO Mentor surgical simulator, all tasks were recorded using Beepa's frames per second screen capture and video recording utility [25] at a rate of 30 frames/s. A digital image-processing algorithm was then designed to track the x and y pixel location of the catheter tip in each frame. The algorithm included a series of segmentation and tracking steps. Segmentation involved converting each image from the RGB to the YCbCr colorspace, establishing a threshold and using a series of morphological operations to isolate the pixels of the catheter from the rest of the pixels in the image. Following segmentation, blob analysis, and spline tracking methods were used to identify the x and y pixel location of the tip of the catheter, per frame.

E. Motor Control Inspired Metrics

Smooth well-coordinated movements are a characteristic feature of well-developed and trained human motor behavior. Movement smoothness in particular has been used as a measure of motor performance of both healthy subjects performing motor control tasks [13], [14], [19], [26] and persons recovering from neurological injury [15]–[17], [27]. The smoothness metrics that typically produce the most significant results are dimensionless, consistent, sensitive to changes in movement, computationally inexpensive, and use most of the available data [26]. In the following sections, a number of methods for computing smoothness that have been used in other domains are presented as potential metrics for evaluating skill level among endovascular surgeons. Each of the motor control inspired metrics is computed from the tool motions captured with either EM sensing or image processing as described in Section II-D. We evaluated non-dimensional jerk, normalized speed, movement arrest period ratio (MAPR), number of peaks in the speed profile, submovement-based metrics, and spectral arc length. These measures are described in detail in the following sections. Key aspects of computation of normalized speed ratio, MAPR, speed peaks, number of submovements, submovement properties, and spectral arc length are illustrated in Fig. 6.

1) Nondimensional Jerk: It has been shown that jerk-based measures with dimensions vary counterintuitively with movement smoothness, whereas a dimensionless jerk-based measure

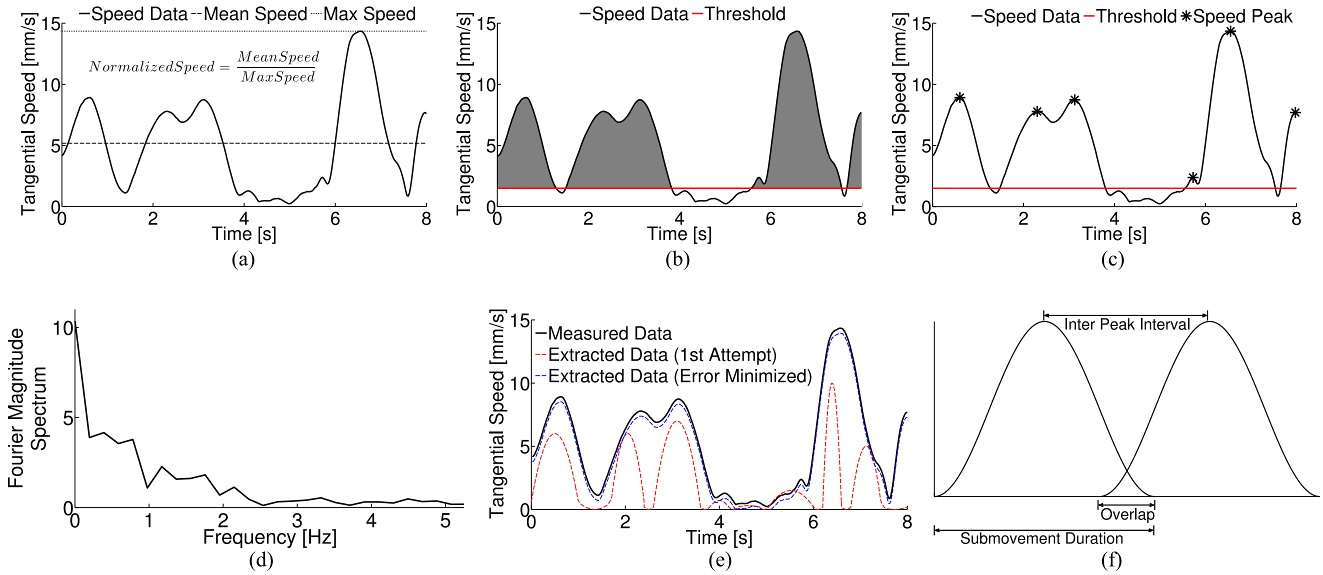


Fig. 6. Visual representation of motor control inspired metrics computed from a representative tangential speed profile. (a) highlights the mean and max speed to compute the normalized speed; the shaded region of (b) represents the percentage of movement when computing the MAPR metric; (c) shows the relevant speed peaks when computing the speed peaks metric; the total length of the Fourier magnitude spectrum shown in (d) would be used to compute the spectral arc length metric; (e) shows a speed profile, along with two extracted profiles using the scattershot algorithm (fully explained in [28]) of computing submovements; the properties associated with submovements are shown in (f).

properly quantifies common deviations from smooth coordinated movement [29]. Performing endovascular procedures involves the control of movements of flexible tools where more experienced surgeons should move the surgical instruments to the desired vessel with less jerkiness than less experienced surgeons. Thus, the dimensionless jerk metric J [29] was used to assess performance while performing endovascular surgical tasks:

$$J = \left(0.5 \int_{T_0}^{T_1} (x'''(t))^2 + (y'''(t))^2 + (z'''(t))^2 \, dt\right) \cdot \frac{T^5}{PL^2} \quad (1)$$

where T_0 and T_1 are the start and end time, respectively, T is the duration, and PL is the 3-D path length.

2) Normalized Speed: For a given motion trajectory, the normalized speed metric is the mean of the speed data divided by the peak speed [16]. The metric's usefulness for skills assessment among healthy subjects is unproven, although it has been used for assessment of persons with stroke [16]. We selected this metric because it was a more effective method of assessing variance in speed than simply computing average speed alone.

3) Movement Arrest Period Ratio: MAPR is the proportion of time that movement speed exceeds a given percentage of peak speed (10% in this study) [30]. A higher MAPR will be computed from a speed profile that tends to have fewer data points near zero speed. While previous studies using MAPR have focused on physically debilitated subjects performing motor tasks [16], [17], the hypothesis in this study is that MAPR can be used in a similar way to differentiate skill level among subjects with different levels of previous experience in performing endovascular tasks.

4) Speed Peaks: The number of peaks in a speed profile represents the number of periods of acceleration and deceleration, such that fewer peaks correspond to smoother movements. They have been used to quantify smoothness in movements of healthy subjects and domain experts [31]–[34] and in stroke patients [16], [35]. In their work, Rohrer *et al.* defined a speed peaks performance metric as the negative of the number of peaks from speed data such that increases in the peaks metric were representative of increases in smoothness [16]. While this is a straightforward method of assessing smoothness of movements, the possible generation of spurious peaks along with possible inconsistent results from speed profiles with a high number of arrest periods means that caution should be taken when drawing conclusions about performance based on this metric.

5) Submovement Extraction: The term "submovements" has been taken to mean the hypothesized discrete units of movement that, when combined, produce a complete movement. Submovements are a theoretically attractive way to define movement because they provide a compact language for concisely coding movement [27]. The existence of submovements has been conjectured to account for human movement when completing both simple point-to-point movements or dynamic pursuit tracking tasks [36], [37]. In the tangential velocity domain, a submovement is represented as a unimodal bell-shaped function. The goal of submovement extraction is to infer the submovement composition of a movement from kinematic data. Rohrer and Hogan [28] proposed an algorithm based on the notion of scattershot optimization to measure the amount of smoothness for a motor control task by extracting the total number of submovements and their associated properties (submovement duration, interpeak interval, and overlap). The scattershot algorithm finds the globally optimal submovement composition probabilistically, i.e., the probability of finding the globally best fit can be made arbitrarily close to 1 by increasing the number of random starting points used in the optimization [28]. In this study, the scattershot algorithm used both support-bounded lognormal (LGNB) curves [38] and minimum jerk profile curves to find the optimal submovement composition. An LGNB curve is defined by [28] according to

$$\begin{aligned} B(t) &= \frac{D\,(T_1 - T_0)}{\sigma\sqrt{2\pi}\,(t - T_0)\,(T_1 - t)} \mathrm{e}^{\left(\frac{-1}{2\sigma^2}\right)\left[ln\left(\frac{t-T_0}{T_1-t}\right)-\mu\right]^2}, \\ T_0 &\leq t \leq T_1, \\ B(t) &= 0, \quad \text{otherwise} \end{aligned} \tag{2}$$

where D is the displacement resulting from the movement, T_0 is the movement start time, and T_1 is the end time. The shape of the support-bounded lognormal curve is influenced by μ (which controls the skewness, or asymmetry, and to some extent the kurtosis, or fatness), and σ (which determines the kurtosis) [28]. Additionally, a minimum jerk profile curve is defined by [27]

$$\begin{aligned} v\,(t) &= \frac{A}{1.875}\left(30\left(\frac{t - t_p + \frac{T}{2}}{T}\right)^2 - 60\left(\frac{t - t_p + \frac{T}{2}}{T}\right)^3 \right. \\ &\qquad \left. + 30\left(\frac{t - t_p + \frac{T}{2}}{T}\right)^4\right), t_p - \frac{T}{2} \leq t \leq t_p + \frac{T}{2} \\ v\,(t) &= 0, \quad \text{otherwise} \end{aligned} \tag{3}$$

where A is the amplitude of the peak, t_p is the time that the peak occurs, and T is the duration of the movement. Using the scattershot algorithm, an increasing number of submovements are fit to each movement until the fit error ε falls below a predetermined threshold (5% in this study). The fit error is defined as [28]

$$\varepsilon = \frac{\int |F(t) - G(t)| dt}{\int |G(t)| dt} \tag{4}$$

where $G(t)$ is the movement speed profile, and $F(t)$ is the extracted speed profile. The literature suggests that as the ability of a subject to perform a task increases, the number of submovements decreases, the submovement duration increases, the extent to which submovements overlap increases and the interpeak interval between submovements descends to an asymptote [27]. In our study, we explore metrics derived from submovements analysis using both LGNB and minimum jerk profile curves as the basis for optimization in order to evaluate differences in the algorithms.

6) Spectral Arc Length: It is intuitive to picture smooth movements as composed of mainly low-frequency components, and unsmooth movements composed of larger amounts of higher frequency components. This inherent association of jerkiness in a movement with higher frequencies implies that the Fourier spectrum of a movement signal could be used to analyze the smoothness of a given movement [26]. One approach to analyzing the frequency spectrum for quantifying smoothness is to look at the complexity of the shape of a Fourier magnitude spectrum. From this idea, the spectral arc-length metric of movement smoothness is defined as negative arc length of the amplitude and frequency-normalized Fourier magnitude spectrum of the speed profile [26]. Using data from a movement with speed profile $v(t), t \in [0, T]$ and duration T, Balasubramanian *et al.*[26] have defined the spectral arc length metric according to

$$\begin{aligned} \eta_{sal} &= -\int_0^{\omega_c} \sqrt{\frac{1}{\omega_c}^2 + \frac{d\hat{V}(\omega)}{d\omega}^2}\, d\omega \\ \hat{V}(\omega) &= \frac{V(\omega)}{V(0)} \end{aligned} \tag{5}$$

where $V(\omega)$ is the Fourier magnitude spectrum of $v(t)$, and $[0, \omega_c]$ is the frequency band occupied by the given movement. $\omega_c = 40\pi$ rad/s (which corresponds to 20 Hz) covers the normal and abnormal aspects of human movements such as tremor [39]. The spectral arc-length metric has been shown to be a valid and consistent measure of movement smoothness [26]. Therefore, this study investigated the metric's utility to assess skill level among endovascular surgeons of various degrees of previous experience.

F. Global Rating Scale for Endovascular Surgery

We aim to identify motor control inspired metrics of performance appropriate for assessing endovascular skill. To this end, we depend on other methods of assessing surgical skill, and explore both caseload and structured assessments to quantify the expertise level of our subjects. We used the global rating scale for endovascular performance (GRSEP), a structured grading tool that is used by a senior clinician to assess each subject based on their performance in endovascular skills, for the latter of these two methods.

The measurable scales of the GRSEP grading tool were: efficiency, ability to manipulate the surgical tools (catheter and guidewire), use of the device, image quality, image safety, autonomy, and level of difficulty, each evaluated on a scale of 1–5. Efficiency of tool movement and ability to manipulate surgical tools are the only scales that assess motor ability rather than procedural knowledge, and as such only these subscales were used to evaluate performance compared with the computed motion metrics. Use of the device assesses proper operation of the simulator system and endovascular tools. Image quality and image safety are scales that assess use of fluoroscopy and are independent of motor skill. Autonomy rates the subject's need for instruction and intervention from the assessor. Level of difficulty assesses the overall difficulty of the task to be performed and was constant for all subjects. Prior to analyzing correlations between the global rating scale and motion-based performance metrics, efficiency and wire/catheter manipulation scores were determined to be the portions of the global rating tool that were most applicable to assessing manual skill rather than procedural knowledge.

In this study, one GRSEP was completed for each session (all subjects completed three sessions; therefore, three GRSEPs were conducted for each subject) as they completed the four endovascular tasks on each apparatus. An average combined score of efficiency plus wire/catheter manipulation (score from 0 to 10) was computed and compared with the quantitative data obtained from the motion capture methods described previously. Additional details about the GRSEP are available in [10].

G. Data Postprocessing and Analysis

Two postprocessing steps were conducted on data obtained from the EM sensors and processed images. In order to eliminate high-frequency background noise, the motion data were filtered with a Savitzky–Golay (S–G) filter using a third-degree polynomial with an 81-frame window. S–G filtering was applied to the measured motion data because of its ability to eliminate most of the noise while preserving the shape qualities of important peaks. After computing the motion-based metrics, both within-subject and between-subject outliers were identified using the 3IQR method. Using this method, 30 within-subject outliers were found. All within-subject outliers were replaced with the individual subject's mean score for that particular metric. No between-subject outliers were identified. Other approaches to outlier replacement are less conservative than our method of replacing with the subject mean, and a number of studies have been conducted to analyze in detail the effects of outlier replacement on statistical analysis of data (see, for example, [40], [41]). To evaluate the implications of our method of outlier removal, we carried out all analyses with outliers in place and compared the results with the output without outliers replaced by the subject's mean. All of the measures yielded comparable output (i.e., all of the candidate metrics showed the same significant correlations, produced significant results for the between-subject ANOVA, and showed significant results for ANOVAs analyzing the within-subject factors). All subsequent analysis is carried out with outliers replaced with subject means.

Data were obtained both from the movement of the catheter tip and the results of the structured global rating scale while subjects executed each of the experimental tasks manually, on both the physical model and simulator. After obtaining the combined global rating score per session for each subject, a Pearson's correlation coefficient was computed to assess the relationship between each measured value from each metric to the global rating scores.

For a subset of metrics, those which were found to correlate with the global rating scale, further analysis was conducted using ANOVA. A mixed design was used, with three within-subject factors (session, platform, and task) and one between-subject factor (ability group) (see Table I). We evaluated between-subject effects and performed posthoc comparisons for significant results to determine the degree to which performance among novice, intermediate, and expert subjects differed. Additionally, we analyzed planned comparisons for a linearly increasing trend for session, where we expected an improvement in performance with each experimental session. We also evaluated and documented within-subject main effects for the platform and task factors. A fully factorial ANOVA model was used; therefore, we could examine interactions as well.

III. Results

We sought to identify performance metrics, computed from motion data, that correlate to expertise as determined by a validated rating scale. Further, we sought metrics that could show significant differences in surgical task performance among domain specialists, and be applicable to tasks performed on multiple platforms. Details on our analysis and findings toward these objectives follow.

TABLE II
Correlation Coefficients (r) and p-Values for Manual Catheterization of Physical Model and Simulator With GRSEP

Metric	Model (r)	Model (p)	Sim (r)	Sim (p)
Jerk	**0.63**	**0.012**	**0.55**	**0.032**
Normalized Speed	−0.05	0.863	0.28	0.321
MAPR	−0.11	0.705	0.21	0.448
Number of Peaks	0.37	0.170	0.37	0.174
Number Submove (LGNB)	**0.80**	**0.001**	**0.71**	**.003**
Number Submove (MinJ)	**0.55**	**0.032**	0.48	0.069
Submove Duration (LGNB)	0.27	0.336	**0.77**	**0.001**
Submove Duration (MinJ)	**0.79**	**0.001**	**0.85**	**0.001**
Inter-Peak Interval (LGNB)	0.09	0.764	−0.11	0.706
Inter-Peak Interval (MinJ)	0.10	0.736	−0.02	0.945
Overlap (LGNB)	0.35	0.196	0.19	0.488
Overlap (MinJ)	0.04	0.887	−0.11	0.708
Spectral Arc Length	**0.77**	**0.001**	**0.84**	**0.001**

Candidate metrics and statistically significant results are indicated in boldface.

A. Critical Failures

Any trials containing critical failures (defined in Section II-C) were either discarded or replaced with the overall subject mean. If a subject had no more than 33% of the trials as critical failures, the data for tasks with critical failures were replaced with the subject's mean metric score; otherwise, the subject was not included in the analysis. In all, we discarded simulator and FEVS data for five out of ten novice subjects. For the remaining 15 subjects (five novice, six intermediate, and four expert), we replaced critical failures with mean metric scores in 18 out of 360 total trials. These replacements were done after the outlier replacement procedure was conducted on the larger set of 480 trials (all 20 subjects). As stated previously, such replacements did not have an effect on the set of metrics that showed significance in terms of correlations or ANOVA analyses.

B. Correlations Between Movement Smoothness and Skill

Table II reports the correlation coefficients for each of the motion-based metrics for manual catheterization on the physical model and simulator. The values in the table were obtained by averaging the data for each metric across all tasks and all sessions, since the correlation coefficients across both of those within-subject factors were not reliably different. An r-value of 0.8 or greater was considered a strong correlation, while a value of r between 0.5 and 0.79 was considered a moderate correlation.

We sought to identify those performance metrics, based on movement smoothness, that most strongly correlated to skill as determined by the GRSEP. As an initial criteria, motion-based metrics were declared candidate metrics if they exhibited more than 33% statistically significant correlations among the

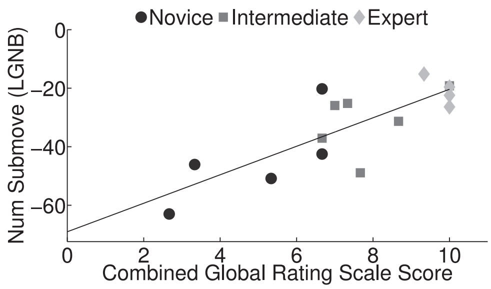


Fig. 7. Scatter plot and regression line for the number of submovements (LGNB profile curves) metric ($\rho = 0.80$, $p = 0.001$). Plot is shown for manual catheterization on the model, averaged across tasks, and sessions.

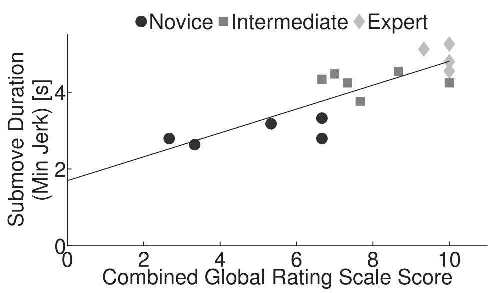


Fig. 8. Scatter plot and regression line for average submovement duration (minimum jerk profile curves) metric ($\rho = 0.85$, $p = 0.001$). Plot is shown for manual catheterization on the simulator, averaged across tasks, and sessions.

twenty-four measured value combinations (e.g., one of two platforms, one of four tasks, and one of three sessions). From the computed results, five candidate metrics were identified:

1) nondimensional jerk;
2) spectral arc length;
3) number of submovements (LGNB profiles);
4) average submovement duration (LGNB profiles);
5) average submovement duration (minimum jerk profiles).

From Table II, it can be seen that all correlation coefficients for every candidate metric except for the average submovement duration (LGNB curves) metric while manually cannulating the physical model were statistically significant. Most of the candidate metrics showed strong correlation with the combined global rating scale score. Figs. 7 and 8 show scatter plots and regression lines for the candidate metric that showed the strongest correlation from tasks performed by manual catheterization on the physical model [number of submovements (LGNB)] and with the simulator [submovement duration (minimum jerk profiles)].

C. Performance Differences Across Ability Groups, Sessions, Platforms, and Tasks

Each candidate metric was further analyzed using ANOVA to highlight statistically significant effects of the between-subject factor of ability group (determined by previous caseload) and of the within-subject factors of session, platform, and task. All reported degrees-of-freedom and p-values were corrected for violating the sphericity assumption. Statistically significant results for ability group were further examined using the Ryan–Einot–Gabriel–Welsch (REGW) posthoc method. Statistically significant session main effects were interpreted using linear contrasts. Contrasts were used to interpret statistically significant interactions among multiple within-subject factors or among the between-subject factor (ability group) and another within-subject factor. Table III shows a summary of the candidate metrics and an "*X*" to represent that a statistically significant result was produced. No higher order interactions were found with the exception of one involving the task factor (see Section III-C4).

1) Differences in Movement Smoothness by Experience Level: Table IV shows the results from our ANOVA analysis of the between-subject ability group factor, evaluated for each of the candidate metrics. The analysis consistently showed a statistically significant difference for the between-subject factor of ability group for all metrics except for nondimensional jerk. Fig. 9 displays the mean metric score for the spectral arc length metric, averaged across all tasks, sessions, and platforms metrics.

The REGW posthoc method was used to test for pairwise differences for the metrics that produced a significant difference for the between-subject factor of ability group. From the REGW analysis, novice scores were found to be significantly different than intermediate and expert scores for the average submovement duration (LGNB) metric while novice, intermediate, and expert scores were each found to be significantly different for the average submovement duration (MinJ), number of submovements (LGNB), and spectral arc length metrics.

2) Longitudinal Effects on Smoothness: Our subjects completed tasks over three separate sessions. We ran planned comparisons for the candidate metrics found in Section III-B using contrasts to evaluate whether a linearly increasing trend in performance existed across sessions, which would indicate learning effects (with alpha adjustments using the false discovery rate method to control Type 1 error rate).

The linear contrast results for the session factor are: For the jerk metric, the linear effect was not significant, [$F(1,12) = 1.29$, $p = 0.278$]; for the number of submovements (LGNB) metric, the linear effect was significant, [$F(1,12) = 12.5$, $p = 0.004$]; for the average submovement duration (LGNB) metric, the linear effect was not significant, [$F(1,12) = 0.09$, $p = 0.774$]; for the average submovement duration (MinJ), the linear effect was significant, [$F(1,12) = 10.29$, $p = 0.008$]; and for the spectral arc length metric, the linear effect was significant, [$F(1,12) = 29.36$, $p < 0.001$]. Fig. 10 displays the mean metric scores for the number of submovements (LGNB) metric, averaged across all tasks, platforms, and ability groups to show differences between session scores for one candidate metric that produced a significant result for the linear contrast on the session factor to illustrate our findings for one representative metric.

3) Differences in Movement Smoothness by Platform: The statistically significant main effect of platform represented a difference in metric scores between manual catheterization on the physical model compared with manual catheterization on the simulator. Fig. 11 shows the platform differences for average submovement duration (computed both with LGNB [$F(1,12) = 8.07$, $p = 0.015$] and Minimum Jerk [$F(1,12) = 177.44$, $p < 0.001$] profiles), the only candidate metric showing

TABLE III
SUMMARY OF STATISTICALLY SIGNIFICANT RESULTS FOR CANDIDATE METRICS

Metric	Correlation With GRESP	Ability Group Differences	Platform Main Effect	Session Main Effect	Platform by Group Interaction	Platform by Session
Jerk	*X*					
No. SubMov	*X*	*X*		*X*		
SubMov Duration (LGNB)	*X*	*X*	*X*		*X*	
SubMov Duration (Min Jerk)	*X*	*X*	*X*	*X*		*X*
Spectral Arc Length	*X*	*X*		*X*		

TABLE IV
ALL CANDIDATE METRICS ANOVA RESULTS FOR BETWEEN-SUBJECT FACTOR OF ABILITY GROUP

Metric	*F*-values and *p*-values
Jerk	$F(2, 12) = 1.23, p = 0.430$
Num Submove (LGNB)	$F(2, 12) = 13.30, p = 0.001$
Submove Duration (LGNB)	$F(2, 12) = 42.50, p = 0.001$
Submove Duration (MinJ)	$F(2, 12) = 134.00, p = 0.001$
Spectral Arc Length	$F(2, 12) = 17.00, p = 0.001$

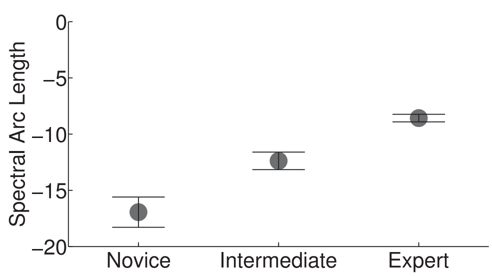


Fig. 9. Mean spectral arc length metric score when performing manual catheterization on the physical model and simulator. Error bars represent standard error of the mean.

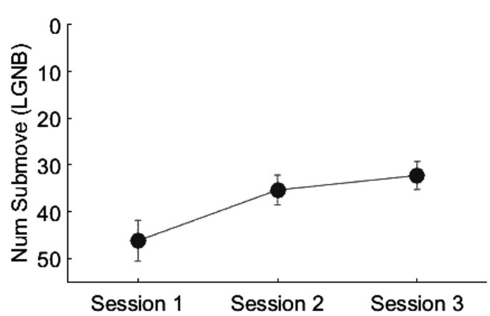


Fig. 10. Reliable session main effect for the number of submovement (LGNB) metric, when performing manual catheterization on the physical model and simulator.

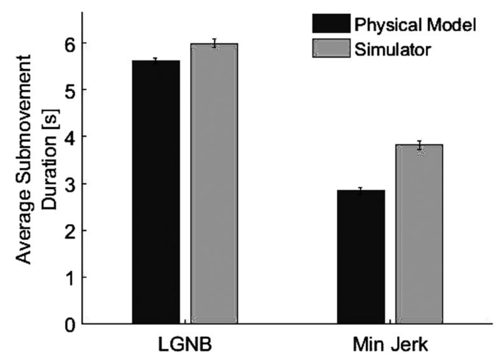


Fig. 11. Statistically significant platform main effects when performing manual catheterization on the physical model and simulator. Error bars represent standard error of the mean.

statistically significant main effects of platform. These results indicate that movements were smoother (as determined by submovement duration) on the simulator than on the physical model.

4) Differences in Movement Smoothness by Task: The main effect of task represented a difference in performance across the four endovascular tasks designed for this experiment. The tasks require navigation of the guidewire and catheter tips to specific locations within the representative anatomy of the physical or simulated model. Because of the varying levels of difficulty in navigating to the targets, completion times and the specific tool movements necessary to access the vasculature varied greatly among subjects. As expected, statistically significant differences were found for the candidate metrics including non-dimensional jerk [$F(3,36) = 4.45$, $p < 0.039$], number of submovements (LGNB) [$F(3,36) = 27.54$, $p < 0.001$], submovement duration (LGNB) [$F(3,36) = 5.65$, $p = 0.03$], submovement duration (minimum jerk) [$F(3,36) = 36.45$, $p < 0.001$], and spectral arc length [$F(3,36) = 19.94$, $p < 0.001$]. Since evaluating the effect of task on performance was not the main focus of this paper, further analysis of the main effects and interactions due to the task factor is omitted. The only higher order interaction that was significant was the task by ability group by platform interaction, and only for the average submovement duration (LGNB and MinJ) metric.

5) Interaction Between Platform and Experience Level: Of the candidate metrics, only submovement duration (LGNB) showed statistically significant results for the platform by ability group interaction (see Fig. 12). An interaction contrast test of the difference in slopes for the two groups was conducted. A Scheffé adjustment was used on the *F*-critical value to control the familywise error rate. The contrast results [$F(1,12) = 35.15$, $p < 0.001$] showed that the trend among different ability groups for the simulator scores was significantly different than the trend of scores when performing tasks on the FEVS model. In other words, the linear trend is significant for increasing submovement duration (LGNB) on the simulator, showing improving performance with increasing experience level. The same trend is not significant for performance on the physical model.

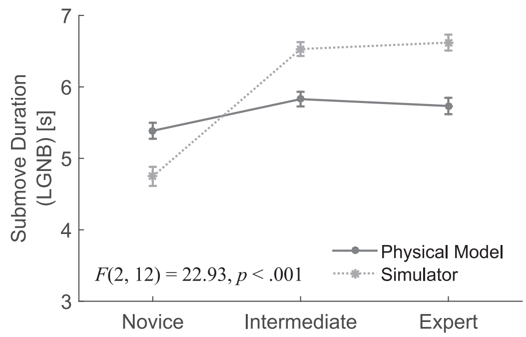


Fig. 12. Statistically significant platform by ability group interaction with F and p-value when performing manual catheterization on the physical model and simulator.

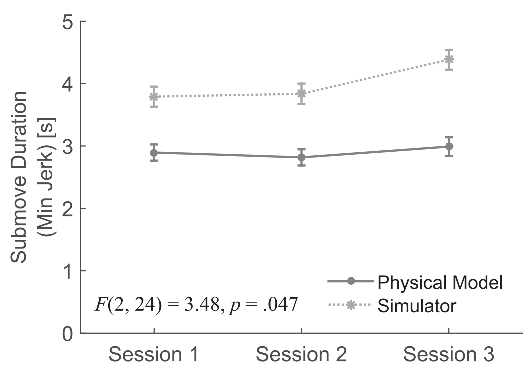


Fig. 13. Statistically significant platform by session interaction with F and p-value when performing manual catheterization on the physical model and simulator.

6) Interaction Between Platform and Session: Next, platform by session interactions were investigated to examine whether any session-to-session learning effects existed among the two platforms. Fig. 13 shows a monotonically increasing trend in submovement duration (minimum jerk profiles) from session-to-session on the simulator, but not when completing tasks on the model. Two contrasts were computed for submovement duration (MinJ), since this was the only metric giving significant results from the interaction analysis. First, the slopes for the session scores were compared across platforms. Second, average submovement duration (MinJ) for Sessions 1 and 2 was compared with submovement duration for Session 3 across platform. A Scheffé adjustment was used on the F-critical value to control the familywise error rate for each contrast that was tested. The results from the contrasts showed that the slopes of the two sets of session scores were different [$F(1,12) = 7.69$, $p = 0.017$] and that the platform effect was different for session 3 than for the averages of session 1 and 2 [$F(1,12) = 8.78$, $p = 0.010$].

IV. Discussion

The teaching of dexterous skill has relied heavily on the apprentice-style training that is built on Halsted's *see one, do one, teach one* approach where surgical residents perform operations under the tutelage of a senior faculty surgeon [2]. This approach to training is limited by a lack of objective and quantitative assessment of skill, instead relying on structured human grading, at best, to evaluate surgical competence. Simulation-based training for endovascular skills has, to date, been focused on rehearsal of procedures and evaluation of outcome metrics such as completion time, fluoroscopy time, and amount of contrast used [42]. Such measures may indeed be useful for assessing skill, but provide little insight into the techniques that surgeons employ.

MIS procedures have altered the way surgeries are performed. One approach to MIS is laparoscopy, which requires superior dexterity and imparts increased cognitive and physical stress on surgeons [43]. Motion analysis of laparoscopic tools has indicated that experts demonstrate superior efficiency of movement when performing procedures [2]. Our study investigated the ability to measure skill level in minimally invasive endovascular procedures using motion based performance metrics inspired by studies of human motor control that indicate smoothness of movements might be linked to task proficiency. We demonstrated that indeed, objective performance metrics that quantified tool movement smoothness for endovascular tasks correlated strongly with expertise as determined by the structured human grading assessments. The motion-based metrics we explored are advantageous compared with structured human grading, since they can be computed without need for observation by known experts, and avoid the potential for subjectivity in using the rating scales. Further, our findings showed that quantitative measures of tool movement smoothness were significantly different among subjects grouped by their previous surgical caseload. Motion-based metrics of smoothness are, therefore, good candidates for assessing surgical skill in an objective and quantitative manner.

Endovascular surgery, compared with laparoscopy, offers an additional challenge of steering flexible instruments. Our work demonstrated that performance metrics derived from tool tip motions in endovascular procedures correlated with surgical skill, despite the added complexity given the flexible nature of the surgical tools. This result is not surprising, given evidence from more fundamental motor control studies that show when motor tasks involve flexible manipulated objects, humans adapt their hand motions from the typically observed minimum-jerk trajectory in order to elicit optimally smooth movements of the object [19].

Our findings support those of similar studies that have reported significant differences in performance of laparoscopic surgical tasks, assessed using motion metrics, on bench-top models [8], [44] and simulators [9], [32]. While most previous studies have shown performance differences between novices with no domain knowledge and domain experts, our smoothness based metrics of spectral arc length, number of submovements (LGNB), and average submovement duration (MinJ) were successful at differentiating skill level among endovascular domain experts (residents, fellows, and attending surgeons). Additionally, this study has demonstrated the successful application of smoothness metrics from nonsurgical domains to endovascular surgical skills assessment. This adds credence to the assertion that movement smoothness is a key characteristic of expert movements and complements findings involving movement smoothness in basic motor control tasks [13], [14] and rehabilitation applications [15]–[17] where smoothness-based metrics were significantly different across ability groups.

A. Correlations With Structured Grading Assessment

We focused our analysis of objective metrics for endovascular skills assessment on motor control inspired metrics that capture movement smoothness associated with skilled movements. Our prior work on developing objective metrics of performance in endovascular tasks showed that motion-based metrics that quantified smoothness and quality of movement were more strongly correlated with the structured grading assessment than those metrics that simply quantified the catheter tool tip kinematics [10]. The findings for metrics that seek to capture movement smoothness are included in this paper for completeness. Specifically, the candidate metrics of spectral arc length, nondimensional jerk, number of submovements (using LGNB curves), and average submovement duration (using both LGNB and minimum jerk profile curves) exhibited statistically significant correlations with the GRSEP scores for manual catheterization. A detailed discussion of the merits of these smoothness metrics, and insights as to why some metrics were more strongly correlated with the structured grading assessment than others, can be found in [10].

B. Ability Group Differences

Number of submovements, submovement duration (MinJ and LGNB), and spectral arc length metrics showed a significant effect of the between-subject factor of ability group, and all metrics showed significant differences between novice surgeons and those with more experience, with the more experienced subjects exhibiting smoother movements when navigating to anatomical targets. This finding is consistent with that of [45], where the tabulated number of movements was shown to be significantly different between inexperienced and experienced surgeons performing laparoscopic cholecystectomy procedures.

C. Learning Effects

Number of submovements, submovement duration (MinJ), and spectral arc length metrics showed a statistically significant linear effect for the session factor, which implied possible learning effects, with scores increasing linearly from session one to session three. For submovement duration (MinJ), there was a significant platform by session interaction, indicating that learning effects were limited to the simulator, and may not be generalizable. These results point to the need for a more focused longitudinal study (either mass training of multiple sessions in a day or distributed training of multiple sessions that span three to four weeks). One should be careful not to extrapolate the results presented here, which suggested learning occurred. It cannot be conclusively stated that actual surgical skill was learned, as opposed to subjects simply becoming more comfortable with the simulator over time. Although our results are consistent with session-to-session learning, whether observed improvements correlate with improvements in actual surgical skill is still an open research question. Indeed, recent findings from robotic laparoscopic procedures show a strong likelihood that device dynamics significantly affect task performance, and that novices should be trained on device operation separate from any task or procedure-specific training unique to a particular surgical domain [18].

D. Effects of Task and Platform

The four tasks chosen for these experiments required subjects to navigate to different locations within the branches of the anatomically inspired FEVS model. While significant differences in performance across task were observed, the effect of task was not central to our study. Our findings indicate that it may be necessary to define a series of tasks to evaluate skill. Indeed, our clinical experts defined more than one task in order to capture a set of navigational skills and challenges for which we could assess performance that could not be captured while navigating to just one single target. Pairwise comparisons of performance across tasks and further diagnosis of interactions involving the task factor are not explored here, but could be investigated in greater detail to ensure that the set of tasks chosen to assess endovascular skills are complete and avoid redundancies.

Only submovement duration, computed with both LGNB and MinJ profiles, showed a significant effect of platform. These results indicate that submovements were longer on the simulator than on the physical model, corresponding to smoother movements on the simulator. It is important to note that submovement duration (LGNB) showed a significant platform by ability group interaction, with novice performance slightly better on the physical model, and intermediate and expert performance slightly better on the simulator. Further, submovement duration (MinJ) showed a significant platform by session interaction, with performance on the physical model remaining fairly constant over the three sessions, while performance on the simulator improved in Session 3 compared with earlier sessions. These results indicate that platform has a modest effect on performance at least in terms of this metric, and learning effects on the simulator may be influencing these findings. Further, one should be careful to draw strong conclusions from platform effects, since the physical model loaded on the simulator has not been extensively validated as equivalent to the real physical model in terms of the realism of tool/vessel interactions. Such studies remain an objective for future work.

E. Recommendations for Endovascular Skill Assessment

In our analysis of performance across ability groups, session, platform, and task, two metrics of smoothness seem best suited to evaluate endovascular surgical skill. Number of submovements, computed using LGNB profiles, was strongly correlated with GRSEP. The between-subject factor of ability group showed a significant effect on the number of submovements, demonstrating significant differences in task performance between novice, intermediate, and expert surgeons. The within-subject factor of session showed a linearly increasing trend in performance based on the number of submovements metric. Our findings, that the number of submovements correlate to skill, are in agreement with other studies that have assessed skill in dexterous tasks by counting movements (similar to computing numbers of submovements) from motion data. Datta *et al.* reported a statistically significant relationship between the experience of a surgeon and the number of hand movements from an experiment involving surgical trainees performing grafting techniques on a vascular model [44]. Another study showed statistically significant differences for a number of movements metric between experienced and inexperienced surgeons performing laparoscopic cholecystectomy procedures [45].

Spectral arc length, a metric that characterizes smoothness in the frequency domain, also showed strong correlation with the GRSEP. The between-subject factor of ability group showed a significant effect on spectral arc length and with this metric, novice, intermediate, and expert performance could be differentiated from each other. The within-subject factor of session showed a significant linear effect. This metric has been shown to be sensitive to alterations in motor behavior, robust to measurement noise, and quantifies smoothness in a consistent manner [26].

Submovement duration computed with minimum jerk profiles, obtained from the submovement analysis scattershot algorithm, also shows promise as a metric for endovascular skill assessment. Submovement duration, like number of submovements, correlated strongly with GRSEP on both platforms. Ability group, platform, and session all showed significant main effects on submovement duration (MinJ), and ability group posthocs showed a capability to differentiate all three ability groups from each other. The presence of platform by group interactions leads us to refrain from more strongly endorsing this metric for surgical skill assessment.

This paper has presented a novel approach to assessing manual skill in endovascular surgical procedures. Motion analysis and evaluation of tool tip movement smoothness offer an important new capability that could be integrated in future training programs where objective assessment of performance during endovascular navigational tasks is required.

Acknowledgment

The authors gratefully acknowledge the resources and facilities provided by the Preclinical Catheterization Laboratory, Houston Methodist Hospital Research Institute, and the timely assistance from the Simbionix Corporation and Northern Digital Incorporated.

References

[1] M. Cox, D. M. Irby, R. K. Reznick, and H. MacRae, “Teaching surgical skills-changes in the wind,” *New Engl. J. Med.*, vol. 355, no. 25, pp. 2664–2669, 2006.

[2] C. Reiley, H. C. Lin, D. D. Yuh, and G. D. Hager, “Review of methods for objective surgical skill evaluation,” *Surg. Endosc.*, vol. 25, no. 2, pp. 356–366, 2011.

[3] H. C. Lin, I. Shafran, D. Yuh, and G. D. Hager, “Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions,” *Comput. Aided Surg.*, vol. 11, no. 5, pp. 220–230, 2006.

[4] P. D. Van Hove, G. J. M. Tuijthof, E. G. G. Verdaasdonk, L. P. S. Stassen, and J. Dankelman, “Objective assessment of technical surgical skills,” *Brit. J. Surg.*, vol. 97, no. 7, pp. 972–987, 2010.

[5] T. T. Tsue, J. W. Dugan, and B. Burkey, “Assessment of surgical competency,” *Otolaryngol. Clin. North Amer.*, vol. 40, no. 6, pp. 1237–1259, 2007.

[6] I. Balasundaram, R. Aggarwal, and L. A. Darzi, “British journal of oral and maxillofacial surgery,” *Brit. Med. J.*, vol. 48, no. 8, pp. 598–606, 2010.

[7] J. Rosen, J. D. Brown, M. Barreca, L. Chang, B. Hannaford, and M. Sinanan, “The blue dragon—A system for monitoring the kinematics and the dynamics of endoscopic tools in minimally invasive surgery for objective laparoscopic skill assessment,” in *Proc. 10th Annu. Med. Meets Virtual Reality*, 2002, pp. 412–418.

[8] S. Cotin, N. Stylopoulos, M. Ottensmeyer, P. Neumann, D. Rattner, and S. Dawson, “Metrics for laparoscopic skills trainers: The weakest link!” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2002, pp. 35–43.

[9] R. V. O’toole, R. R. Playter, T. M. Krummel, W. C. Blank, N. H. Cornelius, W. R. Roberts, W. J. Bell, and M. Raibert, “Measuring and developing suturing technique with a virtual reality surgical simulator,” *J. Am. Coll Surg.*, vol. 189, no. 1, pp. 114–127, 1999.

[10] S. J. Estrada, M. K. O’Malley, C. A. Duran, D. G. Schulz, and J. Bismuth, “On the development of objective metrics for surgical skills evaluation based on tool motion,” in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2014, pp. 3144–3149.

[11] M. K. Chmara, “Trendo tracking system. Motion analysis in minimally invasive surgery,” Ph.D. dissertation, Proefschriftmaken.nl, Amsterdam, The Netherlands, 2002.

[12] T. Flash and N. Hogan, “The coordination of arm movements: an experimentally confirmed mathematical model,” *J. Neurosci.*, vol. 5, no. 7, pp. 1688–1703, 1985.

[13] J. C. Huegel, O. Celik, A. Israr, and M. K. O’Malley, “Expertise-based performance measures in a virtual training environment,” *Presence*, vol. 18, no. 6, pp. 449–467, 2009.

[14] M. K. O’Malley, S. N. Purkayastha, N. Howie, and M. D. Byrne, “Identifying successful motor task completion via motion-based performance metrics,” *IEEE Trans. Human Mach. Syst.*, vol. 44, no. 1, pp. 1–7, Feb. 2014.

[15] O. Celik, M. K. O’Malley, C. Boake, H. S. Levin, N. Yozbatiran, and T. A. Reistetter, “Normalized movement quality measures for therapeutic robots strongly correlate with clinical motor impairment measures,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 18, no. 4, pp. 433–444, Aug. 2010.

[16] B. Rohrer, S. Fasoli, H. I. Krebs, R. Hughes, B. Volpe, W. R. Frontera, J. Stein, and N. Hogan, “Movement smoothness changes during stroke recovery,” *J. Neurosci.*, vol. 22, no. 18, pp. 8297–8304, 2002.

[17] A. U. Pehlivan, F. Sergi, A. Erwin, N. Yozbatiran, G. E. Francisco, and M. K. O’Malley, “Design and validation of the ricewrist-s exoskeleton for robotic rehabilitation after incomplete spinal cord injury,” *Robotica*, vol. 32, no. 8, pp. 1415–1431, 2014.

[18] I. Nisky, A. M. Okamura, and M. H. Hsieh, “Effects of robotic manipulators on movements of novices and surgeons,” *Surg. Endosc.*, vol. 28, no. 7, pp. 1–14, 2014.

[19] J. B. Dingwell, C. D. Mah, and F. A. Mussa-Ivaldi, “Experimentally confirmed mathematical model for human control of a non-rigid object,” *J. Neurophysiol.*, vol. 91, no. 3, pp. 1158–1170, 2004.

[20] J. Bismuth, M. A. Donovan, M. K. O’Malley, H. F. El Sayed, J. J. Naoum, D. M. Peden, E. K., and A. B. Lumsden, “Incorporating simulation in vascular surgery education,” *J. Vasc. Surg.*, vol. 52, no. 4, pp. 1072–1080, 2010.

[21] A. O. Castellvi, L. A. Hollett, A. Minhajuddin, D. C. Hogg, S. T. Tesfay, and D. J. Scott, “Maintaining proficiency after fundamentals of laparoscopic surgery training: A 1-year analysis of skill retention for surgery residents,” *Surgery*, vol. 146, no. 2, pp. 387–393, 2009.

[22] G. B. Zelenock, T. S. Huber, L. M. Messina, A. B. Lumsden, and G. L. Moneta, *Mastery of Vascular and Endovascular Surgery, Section I - Basic Considerations and Peri-operative Care*. Baltimore, MD, USA: Lippincott Williams & Wilkins, 2006.

[23] S. Condino, E. Calabro, A. Alberti, S. Parrini, R. Cioni, R. Berchiolli, M. Gesi, V. Ferrari, and M. Ferrari, “Simultaneous tracking of catheters and guidewires: Comparison to standard fluoroscopic guidance for arterial cannulation,” *Eur. J. Vasc. Endovasc. Surg.*, vol. 47, no. 1, pp. 53–60, 2014.

[24] E. Lugez, H. Sadjadi, D. Pichora, R. Ellis, S. Akl, and G. Fichtinger, “Electromagnetic tracking in surgical and interventional environments: usability study,” *Int. J. Comput. Assist Radiol. Surg.*, vol. 10, pp. 253–262, 2015.

[25] Fraps. (2008). Real time video capture and benchmarking. [Online]. Available: http://www.fraps.com/

[26] S. Balasubramanian, A. Melendez-Calderon, and E. Burdet, “A robust and sensitive metric for quantifying movement smoothness,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 8, pp. 2126–2136, Aug. 2012.

[27] B. R. Rohrer, “Evolution of movement smoothness and submovement patterns in persons with stroke,” Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2002.

[28] B. Rohrer and N. Hogan, “Avoiding spurious submovement decompositions II: A scattershot algorithm,” *Biol. Cybern.*, vol. 94, no. 5, pp. 409–414, 2006.

[29] N. Hogan and D. Sternad, “Sensitivity of smoothness measures to movement duration, amplitude, and arrests,” *J. Mot. Behav.*, vol. 41, no. 6, pp. 529–534, 2009.

[30] H. Beppu, M. Suda, and R. Tanaka, “Analysis of cerebellar motor disorders by visually guided elbow tracking movement,” *Brain*, vol. 107, no. 3, pp. 787–809, 1984.

[31] N. Taffinder, S. Smith, J. Mair, R. Russell, and A. Darzi, “Can a computer measure surgical precision? Reliability, validity and feasibility of the ICSAD,” *Surg. Endosc.*, vol. 13, no. Suppl 1, p. 81, 1999.

[32] E. F. Hofstad, C. Vapenstad, M. K. Chmarra, T. Lang, E. Kuhry, and R. Marvik, “A study of psychomotor skills in minimally invasive surgery: What differentiates expert and nonexpert performance,” *Surg. Endosc.*, vol. 27, no. 3, pp. 854–863, 2013.

[33] V. B. Brooks, J. D. Cooke, and J. S. Thomas, “The continuity of movements,” in *Control of Posture and Locomotion*. New York, NY, USA: Springer, 1974, pp. 257–272.

[34] L. Fetters and J. Todd, “Quantitative assessment of infant reaching movements,” *J. Mot. Behav.*, vol. 19, no. 2, pp. 147–166, 1987.

[35] L. E. Kahn, M. L. Zygman, W. Z. Rymer, and D. J. Reinkensmeyer, “Effect of robot-assisted and unassisted exercise on functional reaching in chronic hemiparesis,” in *Proc. 23rd Conf. IEEE Eng. Med. Biol. Soc.*, 2001, pp. 1344–1347.

[36] J. A. Doeringer and N. Hogan, “Serial processing in human movement production,” *Neural Netw.*, vol. 11, no. 7, pp. 1345–1356, 1998.

[37] N. Hogan, J. A. Doeringer, and K. H. Igo, “Arm movement control is both continuous and discrete,” *Cognitive Stud., Bull. Jpn. Cognitive Sci. Soc.*, vol. 6, no. 3, pp. 254–273, 1999.

[38] R. Plamondon, A. M. Alimi, P. Yergeau, and F. Leclerc, “Modeling velocity profiles of rapid movements: A comparative study,” *Bio. Cybern.*, vol. 69, no. 2, pp. 119–128, 1993.

[39] J. H. McAuley, J. C. Rothwell, and C. D. Marsden, “Frequency peaks of tremor, muscle vibration and electromyographic activity at 10 Hz, 20 Hz and 40 Hz during human finger muscle contraction may reflect rhythmicities of central neural firing,” *Exp. Brain Res.*, vol. 114, no. 3, pp. 525–541, 1997.

[40] J. Orr, P. Sackett, and C. Dubois, “Outlier detection and treatment in i/o psychology: A survey of researcher beliefs and an empirical illustration,” *Personnel Psychol.*, vol. 44, no. 3, pp. 473–486, 1991.

[41] J. Osborne and A. Overbay, “The power of outliers (and why researchers should always check for them),” *Practical Assess., Res. Eval.*, vol. 9, no. 6, pp. 1–12, 2004.

[42] R. Aggarwal, S. A. Black, J. R. Hance, A. Darzi, and N. J. W. Cheshire, “Virtual reality simulation training can improve inexperienced surgeons’ endovascular skills,” *Eur. J. Vasc. Endovasc. Surg.*, vol. 31, no. 6, pp. 588–593, 2006.

[43] H. Xin, J. S. Zelek, H. Carnahan *Laparoscopic surgery, perceptual limitations and force: A review. First Canadian Student Conference on Biomedical Computing; Kingston*, Ontario, Canada. 2006. No. 144.

[44] V. Datta, M. Mandalia, S. Mackay, A. Chang, N. Cheshire, and A. Darzi, “Relationship between skill and outcome in the laboratory-based model,” *Surgery*, vol. 131, no. 3, pp. 318–323, 2002.

[45] S. G. T. Smith, J. Torkington, T. J. Brown, N. J. Taffinder, and A. Darzi, “Motion analysis,” *Surg. Endosc.*, vol. 16, no. 4, pp. 640–645, 2002.

Sean Estrada (S’11–M’14) received the B.S. degree in mechanical engineering from Loyola Marymount University, Los Angeles, CA, USA, in 2003, the M.M.E. degree from Southern Methodist University, Dallas, TX, USA, in 2009, and the Ph.D. degree in mechanical engineering from Rice University, Houston, TX, USA, in 2014.

Cassidy Duran received the B.S. degree from Texas Tech University, Lubbock, TX, USA, in 2003, where she graduated *magna cum laude* with Honors. She then received the M.D. degree from the Temple University School of Medicine, Philadelphia, PA, USA, in 2009.

She is currently a Chief Resident in vascular surgery with the Houston Methodist Hospital, Houston, TX, where she has been in training since 2009. She has extensive research experience in simulation, education, and robotics.

Daryl Schulz received his R.T.(R) [Radiographic Technologist (Registered)] in 1987, and he manages the Pre-Clinical Catheterization Laboratory, Houston Methodist Research Institute, Houston, TX, USA. His research interests include endovascular surgery, cardiovascular intervention, electrophysiological cardiology, and cardiovascular training.

Jean Bismuth received the B.A. degree from Vassar College, Poughkeepsie, NY, USA, in 1989, and the M.D. degree from Copenhagen University, Copenhagen, Denmark, in 1999.

He is currently a Vascular Surgeon with Houston Methodist Hospital, Houston, TX, USA. He also holds an academic position as an Associate Professor with Weill Cornell Medical College, New York, NY, USA.

Michael D. Byrne received the B.A. degree in psychology and the B.S. degree in engineering from the University of Michigan, Ann Arbor, MI, USA, in 1991, and the M.S. degree in psychology in 1993, the M.S. degree in computer science in 1995, and the Ph.D. degree in psychology in 1996, all from the Georgia Institute of Technology, Atlanta, GA, USA.

He is currently a Professor of psychology and computer science, Rice University, Houston, TX, USA.

Marcia K. O’Malley (SM’13) received the B.S. degree from Purdue University, West Lafayette, IN, USA, in 1996, and the M.S. and Ph.D. degrees from Vanderbilt University, Nashville, TN, USA, in 1999 and 2001, respectively, all in mechanical engineering.

She is currently a Professor of mechanical engineering and computer science with Rice University, Houston, TX, USA.