Use of a Machine Learning Algorithm to Classify Expertise: Analysis of Hand Motion Patterns During a Simulated Surgical Task

Robert A. Watson, MBChB, FRCS(Eng)

Abstract

Purpose
To test the hypothesis that machine learning algorithms increase the predictive power to classify surgical expertise using surgeons' hand motion patterns.

Method
In 2012 at the University of North Carolina at Chapel Hill, 14 surgical attendings and 10 first- and second-year surgical residents each performed two bench model venous anastomoses. During the simulated tasks, the participants wore an inertial measurement unit on the dorsum of their dominant (right) hand to capture their hand motion patterns. The pattern from each bench model task performed was preprocessed into a symbolic time series and labeled as expert (attending) or novice (resident). The labeled hand motion patterns were processed and used to train a Support Vector Machine (SVM) classification algorithm. The trained algorithm was then tested for discriminative/predictive power against unlabeled (blinded) hand motion patterns from tasks not used in the training. The Lempel–Ziv (LZ) complexity metric was also measured from each hand motion pattern, with an optimal threshold calculated to separately classify the patterns.

Results
The LZ metric classified unlabeled (blinded) hand motion patterns into expert and novice groups with an accuracy of 70% (sensitivity 64%, specificity 80%). The SVM algorithm had an accuracy of 83% (sensitivity 86%, specificity 80%).

Conclusions
The results confirmed the hypothesis. The SVM algorithm increased the predictive power to classify blinded surgical hand motion patterns into expert versus novice groups. With further development, the system used in this study could become a viable tool for low-cost, objective assessment of procedural proficiency in a competency-based curriculum.

Real operative experience with human feedback and assessment in the operating room (OR) should remain the gold standard surgical training modality, but OR experience can be augmented by simulations that deconstruct operations into component tasks. Increasing the training of surgeons in simulated environments is, however, currently constrained by a reimbursement model that rewards faculty assessor time only in the clinical setting. In this study, we attempted to address this challenge by creating a low-cost, automated assessment tool for the simulated environment that could help prepare residents to optimize the training opportunities in the actual OR and reduce the teaching of basic open surgical skills on real patients.[1] We developed a motion tracking device to attach to the surgeon's dominant hand to capture the motion patterns during a simulated surgical task. Low-fidelity bench models are low-cost simulations of surgical technique that can maintain the real haptic experience.[2] Automating the assessment and feedback of bench model simulated tasks would reduce the need for and expense of direct observation by an expert and increase educational efficiency.

Previous surgical hand motion tracking devices were expensive and used the number of hand movements as an assessment metric that was highly correlated to total task time but was not a measure of the quality of the movements.[3–6] These limitations have prevented widespread adoption of such tracking devices. The only motion economy metric in widespread use is total task time, which needs to be used alongside a measure of quality/error such as in the summative assessment of minimally invasive skills using the fundamentals of laparoscopic surgery tool (which requires human proctoring).[7]

Observation has been the main form of assessment in the apprenticeship model of surgical training: Experienced surgeons recognize and assess trainee surgeons' expertise by observing them operate.[8] If expert surgeons can use this pattern recognition, then, in theory, computers programmed with pattern recognition algorithms could also recognize and classify the different patterns of hand movements that separate the expert from the novice surgeon. It may, therefore, be useful to look at hand motion *patterns*, rather than economy of motion, to evaluate the quality of the surgeon's movements.[9] Computer pattern recognition algorithms currently have applications in speech recognition and in image recognition (e.g., handwriting, face recognition).[10] Machine learning is a branch of artificial intelligence, and a popular and powerful nonparametric supervised learning algorithm used in pattern recognition is the Support Vector Machine (SVM).[11] We have previously detected a difference in pattern structure between the hand movements of novice and expert surgeons and quantified this difference by how easily the patterns could be compressed using the Lempel–Ziv (LZ) complexity metric.[12,13] In a prior study, senior surgeons had more complex hand motion patterns during an open surgical task.[12]

Dr. Watson is assistant professor, Department of Surgery, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Correspondence should be addressed to Dr. Watson, 4026 Burnett-Womack Building, Campus Box 7211, Chapel Hill, NC 27599-7211; telephone: (919) 966-8008; e-mail: robert_watson@med.unc.edu.

Acad Med. 2014;89:1163–1167.
First published online May 21, 2014
doi: 10.1097/ACM.0000000000000316

Given that machine learning algorithms can efficiently recognize different pattern structures, we hypothesized that a machine learning algorithm could increase the predictive power to classify surgical expertise using hand motion patterns. The purpose of this study was to test our hypothesis. By proving our hypothesis, we could develop an assessment tool, with calculated specificity and sensitivity, to classify the expertise of a learner performing a specific surgical task and thus identify both competent residents and fellows and those in need of remediation. It would also offer the possibility to develop a device to provide formative feedback to junior residents during deliberate practice in the learner's own time, and potentially at the learner's own pace, without the need for a commitment of significant time and effort by an expert faculty evaluator.

Method

Participants

The University of North Carolina Medical Center institutional review board approved all procedures. In 2012, we contacted all University of North Carolina at Chapel Hill attending surgeons and surgical residents in their first and second postgraduate years (PGYs 1 and 2) via e-mail and asked if they wished to participate in the study. To standardize the data, left-hand dominance was an exclusion criterion. A consent form was attached to the e-mail invitation so that any questions could be answered prior to obtaining consent. Written consent was obtained at the time of participation, and participation was voluntary. All respondents were included in the study.

Task

Each study participant performed two latex bench model end-to-side simulated venous anastomoses using a continuous suture technique with 6/0 Prolene. A precut 20-mm longitudinal venotomy and identical surgical instruments were used by all participants to standardize the task. We told the participants to complete the simulated venous anastomosis but did not provide samples of an ideal anastomosis or instruction regarding the anastomosis, so all end products reflected the participant's concept of best technique and his or her ability to achieve it.

Data collection

The hand motion of the participant's right hand (up/down, forward/back, right/left, yaw, pitch and roll) was recorded at the rate of 20 times per second (20 Hz) while the participant was completing the latex bench model end-to-side simulated venous anastomosis. The hand motion data were acquired using a custom-made, low-cost (< $200) inertial measurement unit and microcontroller worn on the dorsum of the participant's dominant (right) hand. The device used analog LPR530L (pitch and roll) and LY530ALH (yaw) gyroscopes (STMicroelectronics, Geneva, Switzerland), an ADXL335 triple-axis accelerometer (Analog Devices, Inc., Norwood, Massachusetts), and an ATmega328-based microcontroller (Arduino, Italy). The ATmega328 microcontroller obtained the data from the sensors to upload onto the MATLAB software environment (MathWorks, Natick, Massachusetts). Many other digital or analog sensors and other microcontrollers are easily available that could also be used to make a device to replicate this study.

Data analysis

The hand motion signals were analyzed using custom MATLAB software. Each motion pattern sample was converted into a binary symbolic time series. We used a symbolization scheme based on first-order difference in the observed measurements, considering the difference between two measured values at a time interval apart.[14]

The sequences of symbols were used as the input to calculate an LZ complexity score for each anastomosis.[15] The LZ complexity was normalized by a factor $n/\log_{\alpha} n$ (n = sequence length and α = the number of alphabets in the symbolic sequence [$\alpha = 2$ in the binary sequences]). One LZ complexity score was calculated per anastomosis trial. The LZ metric was tested for linear correlation against the samples' paired task times using the Pearson product–moment correlation coefficient. A receiver operating characteristic (ROC) curve was also plotted to illustrate the performance of the LZ metric as a classifier when its discrimination threshold was varied. The ROC curve was created by plotting the fraction of true positives of the positives (sensitivity) against the fraction of false positives of the negatives (1 minus the specificity) at various threshold settings. ROC analysis provided the optimal threshold to maximize the accuracy of the LZ metric when used as a classifier.

The original SVM algorithm was invented by Vladimir N. Vapnik.[11] Feature extraction was used to reduce the dimensionality of the symbolic time series. These samples were then labeled as expert or novice according to the level of training of the participant (attending surgeons were labeled expert while surgical residents were labeled novice) and used to train the SVM classification algorithm. The trained algorithm was then tested for discriminative/predictive power using unlabeled samples (i.e., the algorithm was blind to the participant's level of training) that were not included in the training of the algorithm.

Therefore, given a set of labeled training examples, each marked as belonging to one of two categories (expert or novice), the SVM algorithm built a model that assigned new examples into one category or the other. The SVM algorithm can efficiently perform nonlinear classification using what is called the "kernel trick."[11] The kernel function that was used in this study was a linear kernel, meaning dot product. The SVM constructed a hyperplane in a high-dimensional space, which was used for classification.

In the field of artificial intelligence, a confusion matrix is a specific table layout that allows visualization of the performance of an algorithm; outside artificial intelligence, the confusion matrix is often called the contingency table or the error matrix (see Chart 1 for an example). Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class. A confusion matrix was created for the LZ and for the SVM classifiers.

Results

Fourteen attending surgeons (experts) and 10 surgical residents (novices) volunteered and participated in the study. Each study participant performed two latex bench model end-to-side simulated venous anastomoses. Of the 14 attending surgeons, 4 were women and 10 were men. All 14 attendings were board certified in general surgery; 2 specialized in surgical oncology, 3 were vascular surgeons, 3 were transplant surgeons, 4 were trauma/general surgeons, and 2 were primarily laparoscopic/bariatric surgeons. The 10 surgical residents included 6

Chart 1

The Confusion Matrix/Contingency Table Used to Visualize the Performance of the Two Classification Methods in This Study[a]

		Surgical expertise (as determined by level of training): Expert	Surgical expertise (as determined by level of training): Novice	
Test outcome	Expert	True positive	False positive (type I error)	Positive predictive value = Σ True positive / Σ Test outcome positive
	Novice	False negative (type II error)	True negative	Negative predictive value = Σ True negative / Σ Test outcome negative
		Sensitivity = Σ True positive / Σ Condition positive	Specificity = Σ True negative / Σ Condition negative	Accuracy = Σ True positive + Σ True negative / Σ Test outcome positive + Σ Test outcome negative

[a]Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class.

women and 4 men. Of the 10 residents, 4 were PGY 1 and 6 were PGY 2.

LZ metric

The LZ metric was tested for linear correlation against the paired task times using the Pearson product–moment correlation coefficient. The LZ metric had a weak negative correlation to total task time: Pearson $r = -0.4$.

The ROC curve for the LZ metric was plotted; this illustrated the performance of the LZ metric as a classifier when its discrimination threshold was varied (see Figure 1). The area under the curve (AUC) was 0.76. An AUC of 1.0 would mean that the test could be used to perfectly discriminate between novice and expert cases, whereas an AUC of 0.5 would mean that the diagnostic accuracy of the classifier is equivalent to that which would be obtained by flipping a coin (i.e., random chance).

ROC analysis proved the optimal threshold to maximize the accuracy of the LZ metric when used as a binary classifier. The most cost-effective LZ threshold was calculated as 0.9008 (see Figure 1) which was then used to classify the hand motion patterns into expert and novice groups. A confusion matrix/contingency table allowed visualization of the performance of the LZ classifier (see Chart 2A). The LZ metric threshold classified the hand motion patterns with an accuracy of 70%, with a sensitivity of 64% and specificity of 80%.

SVM classification algorithm

The trained SVM classification algorithm was tested for discriminative/predictive power against unlabeled bench model anastomosis trials that were not included in the training. A confusion matrix/contingency table allowed visualization of the performance of the SVM classifier (see Chart 2B). The SVM classification algorithm classified unlabeled/blinded hand motion patterns into expert and novice groups with an accuracy of 83%, with a sensitivity of 86% and specificity of 80%.

Discussion

This study proved our hypothesis: A machine learning algorithm increased the predictive power to classify surgical expertise using blinded hand motion patterns.

Machine learning algorithms can learn to recognize and classify patterns automatically. In this study, we recorded hand motion patterns during a simulated surgical task and then trained a machine learning algorithm using these hand motion patterns. After training, the algorithm classified the expertise of blinded surgical hand motion patterns into those of experts and novices, and this study has shown proof of concept using the SVM algorithm. This provided an innovative solution to automating assessment of surgical expertise by applying analytic tools

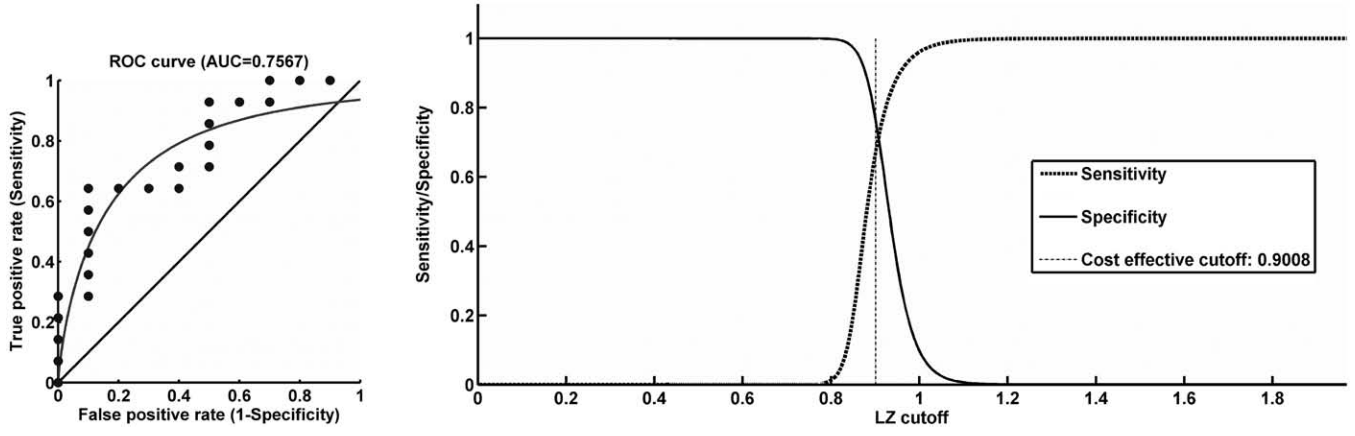

Figure 1 The Lempel–Ziv (LZ) complexity metric as a classifier of expertise in a surgical technique using motion pattern samples captured by a hand motion tracking device on the surgeon's dominant (right) hand, University of North Carolina at Chapel Hill, 2012. Left panel: The receiver operating characteristic (ROC) curve illustrates the performance of the LZ complexity metric as a classifier of surgical expertise when the metric's discrimination threshold is varied. Right panel: The ROC analysis provided the optimal threshold (i.e., the cutoff) to maximize the accuracy of the LZ metric when the metric is used as a binary classifier of surgical expertise.

Chart 2

Performance of the Lempel–Ziv Complexity Metric (Panel A) and the Trained Support Vector Machine (SVM) Algorithm (Panel B) as Classifiers of Expertise in a Surgical Technique Using Hand Motion Pattern Samples From 14 Attending Surgeons (Experts) and 10 Surgical Residents (Novices), University of North Carolina at Chapel Hill, 2012

A. Lempel–Ziv metric		Surgical expertise (as determined by level of training)		
		Expert	Novice	
Test outcome	Expert	9	2	Positive predictive value: 82%
	Novice	5	8	Negative predictive value 61%
		Sensitivity: 64%	Specificity: 80%	Accuracy: 70%

B. SVM classifier		Surgical expertise (as determined by level of training)		
		Expert	Novice	
Test outcome	Expert	12	2	Positive predictive value: 86%
	Novice	2[a]	8	Negative predictive value: 80%
		Sensitivity: 86%	Specificity: 80%	Accuracy: 83%

[a]These two surgeons were the only full-time, practicing laparoscopic attendings.

from the domain of computer science. We challenged the motion economy metrics used in previous surgical hand motion research by developing a new approach that uses hand motion *patterns* to assess open surgical technique and that requires no faculty assessor time.

Unlike motion economy metrics, the binary symbolic time series patterns that we studied do not have a strong correlation to total task time. This may be because the hand motion patterns capture the quality of the surgeon's technique rather than his or her efficiency, although this is conjecture and was not directly evaluated in this study. Although fine finger and instrument manipulations or the tracking of both hands could be expected to give improved data, it is somewhat remarkable that our crude measure of surgical technique (i.e., the motion patterns of a single point on the dorsum of the dominant hand) could enable the algorithm to distinguish a significant difference between novice and expert surgeons. Our method has the real advantage of low cost and does not require (expensive) human expert assessors.

This study does have significant limitations. It was task specific and conducted only in a simulated environment, outside the OR. The study sample size was small and from a single institution, and the participating faculty and residents were volunteers rather than a random sample. However, our quantitative technique showed significant results, and increasing the training sample size could improve the performance of the machine learning classifier if overfitting were avoided. We did not examine the quality of the finished anastomoses and have not proved that the classifier is an indicator of the quality of task outcome. We did not evaluate real surgical skill, and we assumed that attending surgeons had more surgical skill and residents had less; therefore, we only demonstrated surrogate construct validity where level of training was a surrogate marker of real surgical skill/performance. However, this limitation led to the interesting finding that the only two false negatives that were classified by the SVM (attending surgeons labeled as novice) were the full-time practicing laparoscopic surgeons, who had no routine day-to-day experience using an open technique of vascular anastomosis, possibly bringing into question the maintenance of this competency in these two surgeons. We cannot speculate from our data if significant differences in individual surgeon hand motion have an effect on final product outcome, as that was not the objective of this study. We have only proved that hand motion patterns can predict the level of expertise of a participant, using a participant's grade (attending versus resident) as the marker of competency between cohorts. In the future, any correlation between quality of final product and the motion patterns should also be studied.

The use of machine learning software is a new avenue of research in surgical education. We believe that the expansion and development of the methods we used could form the basis of low-cost educational tools to evaluate procedural proficiency and increase educational efficiency to ultimately improve patient safety. With further development, our system could become a viable tool for objective assessment in a competency-based curriculum. Ultimately, artificial intelligence and machine learning techniques hold potential for monitoring surgeons' performance. Future technology, more sophisticated than that used in this study, could be used routinely on practicing surgeons in the OR with direct application to group quality assurance activities and/or the early detection of impairment due to the effects of aging and illness.

Acknowledgments: The author greatly appreciated the participation of the faculty and residents of the University of North Carolina at Chapel Hill who took part in this study.

Funding/Support: This study was supported by a University of North Carolina at Chapel Hill Academy of Educators award.

Other disclosures: None reported.

Ethical approval: The University of North Carolina Medical Center institutional review board approved all procedures.

References

1 Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: A systematic review and meta-analysis. JAMA. 2011;306:978–988.
2 Norman G, Dore K, Grierson L. The minimal relationship between simulation fidelity and transfer of learning. Med Educ. 2012;46:636–647.
3 Brydges R, Sidhu R, Park J, Dubrowski A. Construct validity of computer-assisted assessment: Quantification of movement processes during a vascular anastomosis on a live porcine model. Am J Surg. 2007;193:523–529.
4 Mackay S, Datta V, Mandalia M, Bassett P, Darzi A. Electromagnetic motion analysis in the assessment of surgical skill: Relationship

between time and movement. ANZ J Surg. 2002;72:632–634.
5 Datta V, Bann S, Mandalia M, Darzi A. The surgical efficiency score: A feasible, reliable, and valid method of skills assessment. Am J Surg. 2006;192:372–378.
6 Memon MA, Brigden D, Subramanya MS, Memon B. Assessing the surgeon's technical skills: Analysis of the available tools. Acad Med. 2010;85:869–880.
7 Fried GM, Feldman LS, Vassiliou MC, et al. Proving the value of simulation in laparoscopic surgery. Ann Surg. 2004;240:518–525.
8 Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): A tool to assess surgical competence. Acad Med. 2012;87:1401–1407.
9 Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. Surg Endosc. 2011;25:356–366.
10 Jain AK, Duin RPW, Mao J. Statistical pattern recognition: A review. IEEE Pattern Anal Mach Intell. 2000;1:4–37.
11 Vapnik V. The Nature of Statistical Learning Theory and Application. New York, NY: John Wiley Publishing; 1995.
12 Watson RA. Quantification of surgical technique using an inertial measurement unit. Simul Healthc. 2013;8:162–165.
13 Watson RA. Computer-aided feedback of surgical knot tying using optical tracking. J Surg Educ. 2012;69:306–310.
14 Kurths J, Schwarz U, Witt A, Krampe R Th, Abel M. Measures of complexity in signal analysis. In: Chaotic Fractal Nonlinear Signal Process. 3rd Technical Conference on Nonlinear Dynamics and Full Spectrum Processing; Mystic, CT; July 1995. AIP Conference Proceedings 375. Woodbury, NY: American Institute of Physics; 1996:33–54.
15 Kaspar F, Schuster HG. Easily calculable measure for the complexity of spatiotemporal patterns. Phys Rev A. 1987;36:842–848.

Teaching and Learning Moments
Reality Check: A Reflection on Refugee Health

Groggy medical students and a hematologist gathered around the table on a Monday morning for a tutorial session. We were there to discuss the case of Iman, a seven-year-old child who came to the clinic for a routine examination. He and his family had recently immigrated to Canada as refugees. As my colleagues and I began the discussion, we scrutinized lab results, which showed microcytic anemia. Our group discussed the differential diagnosis, then we collectively decided to order iron studies and a hemoglobin electrophoresis. The tests uncovered what we suspected, and a diagnosis of beta thalassemia minor was given to Iman. Ours were brief exchanges about different kinds of thalassemias, treatment options, and side effects. We asked questions and answered them. It seemed like a relatively simple exercise.

As our time was winding down, we considered what impact our diagnosis would have on Iman and his family. I paused for a moment and scanned the patient description. Iman's refugee status grabbed my attention. "I don't think Iman would even have been diagnosed," I blurted out. Some of my colleagues gave me puzzled glances. I explained: "I don't think he would have received care under current Canadian regulations because many refugees are not eligible for routine medical examinations. The government recently made significant cuts to refugee health care."

Some of my colleagues understood what I was referring to and some looked surprised. The recent cuts prevented some refugees from receiving basic health services, like prenatal screening and routine medical examinations. Without health insurance coverage, many refugees were discouraged from seeking proper health care or were turned away at the clinic. My colleagues and I spent the last few minutes of the session talking about disparities in access to health care.

Reflecting on this session, I realized that my colleagues and I had spent almost an entire hour talking about history taking, physical examination, lab tests, and treatment options, when in reality, Iman and his family would probably not even have come into the clinic because they lacked health insurance coverage. Reaching a correct diagnosis and discussing comprehensive treatment options were irrelevant if Iman and many other refugees in similar situations could not access the health care that they needed. Perhaps we should have started the session with a discussion about Iman's refugee status and how that affected his access to health care and treatment options.

From this experience, I gleaned that diagnosis and treatment of disease cannot be separated from the social context of our patients' lives. In addition to the scientific evidence and clinical principles that we need to consider, we must not forget to look at the whole patient and consider how social context can impact health. Moreover, I realized that understanding the social determinants of health can provide valuable information to meet our patients' unique health needs. Educating ourselves about our patients' health insurance coverage, access to health care, and policy changes are just as important as learning about the underlying causes and management of disease. Finally, being a good physician means not only learning about our patients' social context but also advocating for changes—from the clinic to the community level—that will allow our patients to access quality health care, regardless of where they are from.

Author's Note: The name in this essay has been changed to protect the identity of the patient.

Matthew J. To, BMSc

Mr. To is a medical student, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada; e-mail: mto3@alumni.uwo.ca.