Received: 19 June 2017 | Revised: 22 September 2017 | Accepted: 25 September 2017

DOI: 10.1002/rcs.1873

WILEY The International Journal of Medical Robotics and Computer Assisted Surgery

ORIGINAL ARTICLE

Performance metrics for guidance active constraints in surgical robotics

Nima Enayati | Giancarlo Ferrigno | Elena De Momi

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

Correspondence
Nima Enayati, Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133, Milan, Italy.
Email: nima.enayati@polimi.it

Funding information
European Union's Horizon 2020 Industrial Leadership, Grant/Award Number: H2020-ICT-26-2016-1-732515

Abstract

Active constraint (AC)/virtual fixture (VF) is among the most popular approaches towards the shared execution of subtasks by the surgeon and robotic systems. As more possibilities appear for the implementation of ACs in surgical scenarios, the need to introduce methods that guarantee a safe and intuitive user-interaction increases. The presence of the human in the loop adds a layer of interactivity and adaptability that renders the assessment of such methods non-trivial. In most works, guidance ACs have been evaluated mainly in terms of enhancement of accuracy and completion time with little regard for other aspects such as human factors, even though the continuous engagement of these methods can considerably degrade the user experience. This paper proposes a set of performance metrics and considerations that can help evaluate guidance ACs with reference to accuracy enhancement, force characteristics and subjective aspects. The use of these metrics is demonstrated through two sets of experiments on 12 surgeons and 6 inexperienced users.

KEYWORDS

computer assisted surgery, human-machine interfaces, telesurgery, haptics

1 | INTRODUCTION

While current practice in surgical robotics is limited to direct human control, efforts are being made towards autonomous task execution that describe a future in which the clinicians only supervise most of the procedures performed in the operating room.[1,2] However, for such a transition to happen, robotic systems must demonstrate a substantial level of intelligence, cognition and reliability that may not be achieved soon. Current advances in medical robotics research have motivated researchers to investigate cooperative control methods in which humans and robotic agents collaborate to achieve a shared goal.[3] Surgery is an interweave of many basic procedures into a cognitively demanding activity, and techniques such as Active Constraints (AC) (also known as virtual fixtures) have the potential to significantly enhance the synergy between the clinicians and their instruments in conducting these basic procedures. Even though ACs have been employed in clinical procedures that involve interactions with rigid bones,[4] the complexities such as soft tissue localization, registration and tracking[5] have precluded the widespread use of these assistive methods. In addition, implementing ACs for surgical applications faces challenges such as control loop stability that have prevented today's commercial teleoperated robotic systems from having haptic feedback.[6] Nonetheless, with the rapid progress in computer vision, robotics and artificial intelligence technologies, one can expect to see robotic systems play a greater part in physical interactions in surgery and share the control of the instruments with the surgeons.

The implementation of active constraints generally involves two main phases.[7] In the first phase, the geometrical properties of the constraint are defined by either a user or an autonomous agent. The defined geometry may be modified during the task for reasons such as accounting for tissue deformation or displacement. In the second phase, the assistive force is generated based on information about the motion of the surgical tool (i.e. pose, velocity, etc.) relative to the constraint's geometry. The definition of geometry and the tracking are inherently task-dependent, and methods can be evaluated by comparing the generated constraint shape with some corresponding ground truth. In contrast, the methods of force generation are applicable to wide ranges of tasks, and evaluation of their performance is not trivial.

Generation of AC geometry has been the focus of numerous works in the literature because tissue deformability and the lack of structural features create major challenges in defining and updating the geometries. Examples include concave tunnels,[8] dynamic admittance-type,[9] streaming Point Clouds-based[10] and 3D eye gaze-based.[11] Fewer studies have proposed enforcement methods,[12-14] but as the generation of

geometry and methods of tracking improve and practical implementations become more likely, the need for novel and efficient enforcement methods will grow. Studies on ACs have used a few metrics to prove the efficacy of their methods in enhancing accuracy and safety.[13,15,17] However, the discussions are often limited to these aspects. Occasionally, subjective aspects have been taken into consideration. For instance, the cognitive demand imposed on the user by the proposed assistive method was measured[11] through functional near-infrared spectroscopy (fNIRS). Most works, however, have not provided extended discussions on subjective aspects. There is no doubt that ACs are primarily designed to improve accuracy and safety. Nevertheless, if these methods are to be developed for future clinical implementation, their attributes in interactive use and the engendered subjective perceptions must be thoroughly studied. Surgeons are highly sensitive to any medium that may interfere with their interaction with the tissue. If an assistive system causes unexpected or unintuitive motions, surgeons may show reluctance in using it, no matter how greatly it improves the accuracy of the surgical task. There is a need to formalize measures that can be used to reliably evaluate the performance of guidance active constraints considering all involved aspects, especially for works that propose enforcement methods.

This work discusses factors that should be considered in the design and development of guidance ACs, specifically accuracy enhancement, force quality and subjective aspects. Regarding the accuracy metrics, those used in the literature are revisited and their significance is discussed. Subsequently, force assessment metrics are proposed that can extract and depict the features of the generated AC forces quantitatively, using time-domain and frequency-domain analyses. The last set of proposed metrics focus on subjective aspects such as time-efficiency and cognitive ease. The use of the metrics is then further demonstrated by applying them to the experimental results acquired from 12 surgeons and 6 non-experts performing tasks using 4 guidance AC methods in a virtual environment. The C++ codes of our implementation of these methods are available online at:

https://github.com/nearmrs/dynamic_active_constraints for interested readers. It should be noted that the term 'active constraint' is used here to represent assistive methods that explicitly affect the motion of the surgical tool, and visual or auditory feedbacks are not considered. Furthermore, the focus of this work is on the more popular category of impedance-based active constraints that generate guidance forces/torques based on motion measurements.

2 | ANALYSIS OF ACCURACY ENHANCEMENT

The aim of active constraints in robot-assisted surgical applications is to enhance accuracy and safety by constraining the motion (in terms of position, orientation, velocity, etc.) of the surgical tool. Constraints are built upon a defined geometrical shape (point, line, area, volume, etc.), from which deviations are undesired. In the case of forbidden region ACs that are often areas or volumes, performance enhancement can be estimated by counting the number of incidents of penetration into the forbidden region and the length of penetrations. These incidents are more of a discrete nature, and depending on the level of difficulty of the task, they may occur occasionally. On the other hand, for guidance ACs the deviation from the defined shape (usually curved lines) must be measured continuously at the system's rate, resulting in a time-series of spatial errors. The deviation is typically calculated only when the tool is in contact with the tissue. Various statistical representations can be used to interpret the tracking error time-series acquired in experimentations. Works in the literature have used the average,[15-18] average and standard deviation,[16] RMS (root mean square)[13] and maximum value[17] of the tracking error time-series. In the following, some statistical measures that can provide insight into the performance in accuracy enhancement of active constraints are explored. It must be noted that, as most of these measures condense a large amount of information into a single value, they emphasize only certain aspects of the characteristics of accuracy enhancement of the ACs. Therefore, providing multiple metrics is generally preferred in order to draw a more complete picture of the error distribution.

2.1 | Error distribution

Although statistical representations are needed to quantitatively evaluate the performance of AC methods, graphical representation of the tracking error distribution can be a good starting point in the analysis. Histograms give a rough sense of the density of the underlying distribution of the error and allow a quick assessment of the performance in a qualitative fashion. When comparing different acquisitions, it can be more convenient to use a uniform bin width and normalize the histograms so that the bar heights add to 1 (Figure 1).

2.2 | Mean, median, RMS and IQR

The mean error (ME) and the mean absolute error (MAE) are widely used measures in studies of accuracy enhancement. Assuming that errors in all directions are equally undesired, the tracking error defined as the distance from the closest point on the desired geometry will

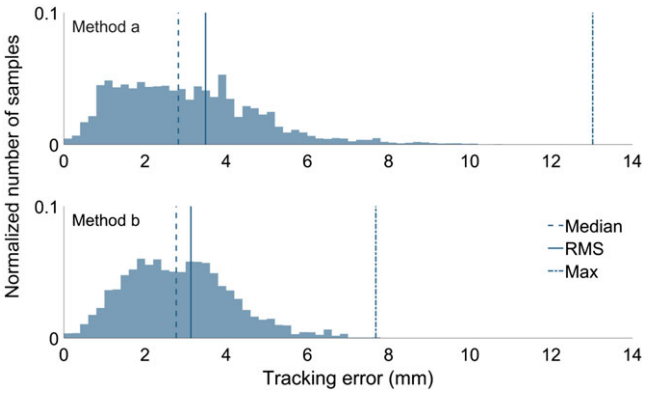

FIGURE 1 Normalized histograms of the tracking error acquired from a task performed using two different guidance active constraints. While the median errors are almost identical (3.0 and 2.9 mm for methods (a) and (b) respectively), the histogram shows that method (b) has lower frequencies at larger error values. This difference is captured by the RMS to some extent (3.5 and 3.1 mm respectively), since it penalizes larger errors more. The maximum error, however, shows a significant difference

always be a value larger than or equal to zero. If different directions are considered and the error includes negative values, MAE is preferred over ME as it prevents error cancelation. The RMS error (RMSE) or RMS deviation (RMSD) is a frequently used measure of the pairwise difference between values estimated or predicted by a model and the actual values, which can serve as a measure of how far on average the error is from zero. RMSE differs from the mean of the pairwise differences in that the latter does not measure the variability of the difference, and the variability as expressed by the standard deviation is around the mean instead of zero. While the MAE gives the same weight to all errors, the RMSE penalizes variance because it gives more weight to errors with larger absolute values. When both metrics are calculated, the RMSE is by definition never smaller than the MAE. The MAE is suitable to describe uniformly distributed errors. In model error analysis it has been argued the RMSE has an advantage over the MAE in illustrating the error distribution because model errors are likely to have a normal distribution rather than a uniform one.[19] Although the error distributions in tasks such as path following are not necessarily Gaussian (and based on our experience they rarely are), the penalization of larger errors can be appropriate for surgical application where large errors may cause safety hazards. Nevertheless, mean errors are particularly susceptible to the influence of outliers and skewed data, and it is known that non-normal populations are better represented by the median. Thus, median and RMSE are generally more appropriate measures of the central tendency of tracking error for ACs.

Active constraints can help surgeons to maintain a more constant motion with lower variations in error.[20] Such enhancements in precision, however small they seem, may be beneficial for operations requiring sub-millimetre accuracy. For non-normal error distributions, interquartile range (IQR) can be used to quantify the amount of dispersion of the population. The acquisitions depicted in Figure 1 provide a good example of the necessity of providing dispersion metrics along with measures of central tendency. As can be seen, the medians of the methods are very close: 3.0 mm for method (a) and 2.9 for method (b). The IQRs of the acquisitions (2.3 and 1.7 mm respectively for (a) and (b)), however, show a higher dispersion for method (a), suggesting a better performance in variability reduction for the latter method, which can also be observed from the histograms.

2.3 | Maximum

In tasks where large errors may cause serious complications, the extreme aspects of the error population are of more interest than the accuracy metrics of the central tendency. The maximum tracking errors depicted in Figure 1 show that two error populations with comparable medians may have significantly different maxima.

In some cases, there may be a certain threshold above which the error raises concern. Here, the number of deviations larger than the threshold can be used as a metric of accuracy enhancement along with the maximum deviation.[14]

2.4 | Intersubject performance variability

The discussed metrics can also be computed per subject to analyze the variation of performance among users for each guidance method. Guidance ACs may reduce the dependency of the accuracy performance on the user's skills to some extent. Studying the intersubject performance variability can reveal effects of subjects' skills on the performance and the ability of the guidance AC to reduce the variability in the results.

3 | FORCE ANALYSIS

Since guidance ACs aim to assist clinicians by attempting to redirect the motion of their hands, it is crucial to design them such that operators find them natural. The surgeons who participated in this work's experiments described an ideal assistive force as least disturbing, gentle, intuitive and fluid. To quantify such adjectives and inspect the generated forces, it is necessary to use a set of measures that capture the desired characteristics. However, to our best knowledge, no work on active constraints/virtual fixtures in the literature has used a form of metrics to investigate the performance of the forces generated by the guidance methods. We propose a few metrics and considerations in time and frequency domains that can be employed to evaluate the characteristics of forces generated by guidance ACs quantitatively. These metrics are applicable to enforcement methods of static and dynamic types regardless of the point of application of the AC (only tool tip or more points).

When evaluating guidance ACs, in addition to the magnitude of the generated force, the changes in its direction can carry useful information. To analyze both magnitude and direction using the minimum number of components, Cartesian force vector (f_x, f_y, f_z) can be converted to spherical coordinate system, where the radial distance represents the magnitude of the force, f, and the azimuth θ and polar φ angles represent the directional components of the force vector.

3.1 | Time-domain analysis

In force analysis, the characteristics of the generated force's progression in time are more informative than statistical representations such as those used in accuracy analysis. For instance, the maximum applied force is usually set in the design of the active constraint enforcement and is already known. Nevertheless, measures of the central tendency of the magnitude of forces generated by the AC in a task can be employed as a generic metric of its activity. For example, a higher mean value of force magnitude may suggest that the AC has been more engaged. As will be discussed in the following, investigating the variability of the force in terms of magnitude and direction can lead to better understating of the qualitative characteristics of an AC.

3.1.1 | Force derivative

Higher rates of change of magnitude or direction may lead to abrupt engagements and are generally undesired. To study the variability of the generated force, the derivative of its magnitude and direction components with respect to time can be used. Since the force applied on a constant mass is proportional to acceleration, the time derivative of forces applied on a constant mass is proportional to the derivative of the mass's acceleration, i.e. jerk. A good everyday example of the effects of jerk is changing gears while driving a car: although the engine

power limits the accelerating force, an inexperienced driver can experience severe jerk, because of intermittent force engagement over the clutch. Minimization of jerk has been studied in many engineering fields, in particular in applications that encompass machines interacting with humans. Studies of coordination of voluntary human arm movements have shown that a major goal of human motor coordination is the production of the smoothest possible movement of the hand that can be modelled by a trajectory with minimum-jerk time-profile.[21] Thus, limiting the jerk can enhance the intuitiveness and smoothness of the forces applied to the operator's hand. In fact, in the field of assistive robotics for rehabilitation, where robotic mechanisms assist stroke survivors in performing movements with their impaired limb, generating minimum-jerk trajectories is a ubiquitous practice.[22]

3.1.2 | Angular rate

The derivative of the azimuth and polar angles can reveal how continuously the direction of the force changes. Large and fast variation in the force direction may increase the chance of bouncing on boundaries and creating vibrations. However, angular rate can be considered as a secondary metric with respect to the jerk. The reason is that the distraction created by high angular rates depends also on the magnitude of the force. In other words, if at an instant, the angular velocities are high but the corresponding force magnitudes are negligible, no disturbance may be felt by the user. Therefore, a better metric would be one that takes into account the magnitude of the force too. This can be achieved by scaling the angular rate at each instant by the normalized magnitude of the force vector. The scaled angular rates can be calculated as:

$$\dot{\varphi}_s = \frac{f}{F_{\max}}\dot{\varphi}, \qquad \dot{\theta}_s = \frac{f}{F_{\max}}\dot{\theta} \tag{1}$$

where $F_{\max}$ is the maximum magnitude of force generated by the AC.

3.2 | Frequency-domain analysis

As was mentioned before, active constraints with highly fluctuating force can be disturbing and are undesired. The frequency domain is particularly apt for studying variability of the force. The power spectral density (PSD) of AC force data acquired during execution of a task can reveal some aspects of the AC behaviour. However, since the generated force depends on the motion and the motion generally does not follow a particular pattern, the power spectrum can contain a wide range of frequencies, making it less illustrative. In particular, in cases where the performances in force generation of two AC methods are to be compared, it can be helpful to design experiments that comprise simple tasks and well-defined motions to facilitate the extraction of information from the force power spectra.

If estimated directly from a fast Fourier transform of the signal, the spectrum may contain a large amount of noise. Welch's method is a popular approach to estimate the PSD with reduced noise through averaging periodograms across time. Since the force signal is real-valued, a one-sided PSD estimate can be used. Ideally, the power should decrease from the DC frequency as the frequency increases. Evident peaks in the spectra can indicate the presence of rapidly oscillating forces that may cause undesired behaviours. It should be noted that, although the discussed metrics are regarding forces generated by an AC, a similar concept applies to generated torques.

4 | SUBJECTIVE ASPECTS

An indispensable phase in design and development of new tools is the subjective evaluation through rigorous tests on targeted users. Subjective evaluation focuses on the user's perception, feelings and preference, and can reveal issues potentially overlooked in objective assessments. Haptic guidance imposes sensory/control transformations on the human operator, which are the key factor in determining a subjective performance evaluation of the method. A popular approach to estimating subjective performance is to directly ask users about their experiences after the tests—that is, subjecting them to the system and then administering questionnaires about their personal opinion and feelings.[23] The questions can concern general aspects such as whether the user will be willing to employ the tested method in real practice in the future, or which one of the tested methods they prefer. In contrast, specific questions (e.g. to rate the distraction level) can shed light on certain aspects of the performance of the guidance method.

In a recent study, Koehn and Kuchenbecker[24] conducted two experiments on human subjects to discover whether surgeons and non-surgeons value the addition of vibration feedback from surgical instruments during robotic surgery. This work presents a good example of general and application-oriented subjective assessment. After performing the experiments with and without feedback, the subjects answered a survey by rating (from 0 to 100, corresponding to 'strongly disagree' to 'strongly agree') statements such as 'the haptic feedback caused peculiar or undesirable sensations in my hands' or 'the haptic feedback made me more aware of my instrument contacts'. The results show that most subjects in both experiments (95% and 98%, respectively) preferred receiving vibration feedback from tools and that all subjects in the second experiment agreed 'it would be useful for surgeons to have the option of such feedback'.

Sole reliance on questionnaire results, however, cannot verify the efficacy of methods. It can be argued that questionnaire-based measures after the event cannot in principle rule out the possibility that the reported quality was called into being simply by its having been asked about.[25] Questionnaires can be most useful in circumstances where the subjects are provided with a stock of experience against which to judge a given experience and make comparisons. Thus the experiments and corresponding questions must be designed having in mind that the outcomes rely on respondents being able to reasonably compare a given situation with a number of other situations, and make a quantitative assessment. Nevertheless, such an evaluation must be considered as a complementary analysis and the interpretation of its results should ratify (at least to some degree) similar findings from objective assessments.

Since humans attribute their own meaning to qualities and perceptions, the statements and questions have to be precise and clear to minimize the probability of unintended attributions. Furthermore, it must be certified that the user's performance with a method does not affect their answers when asked about other aspects of that method. The level

of expertise of the subjects should also be taken into consideration in the design of surveys. While non-clinician participants can be studied for human perceptions and general cognitive aspects, surgeons can be asked about specific aspects and more application-oriented concerns. The inter-rater reliability of agreement can be statistically estimated using Fleiss' kappa measure.[26] It can be interpreted as the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings randomly. Although guidelines have been given for interpreting the kappa values, there is no generally agreed-upon measure of significance.

Another subjective aspect to be considered is that of cognitive load, which indicates the total amount of mental effort being expended by the subject to perform a task. Functional near-infrared spectroscopy (fNIRS) is an optical neuroimaging technique that could potentially reveal the cognitive demand involved in performing a task by detecting changes in brain haemodynamics that are reflective of the brain's function. Using fNIRS, it is shown[11] that when a novel haptic constraint is applied, there is an evident difference in activating the prefrontal cortex between expert users and novices. Interestingly, the estimated cognitive demand on the novice users was found to be higher when the assistive method was activated than with free acquisitions. Such assessments can provide valuable insights into user experience and the learning curve of an assistive method. However, the complexities and uncertainties involved in explicit measurements of cognitive load have led researches to rely on measures such as temporal demand as an implicit manifestation of the efficiency and coordination with which actions are performed during a surgical task. Assistive methods that effectively reduce the cognitive load are believed to decrease the overall time subjects require to execute a task successfully. For the study of subjective aspects and workload including the psychological aspects, the NASA Task Load Index (TLX)[27] can be used. This well-known framework proposes a human-centred definition of workload consisting of six metrics: mental, physical and temporal demands, frustration, effort and performance. It has been assumed that some combination of these metrics IS likely to represent the workload experienced by most people performing most tasks.[28]

5 | EXPERIMENTS

A set of experiments were performed to demonstrate the application of the proposed metrics. These experiments are described in the following and the results are analyzed in order to compare various aspects of the performance of three guidance active constraints. The experiments were carried out in accordance with the recommendations of our institution with written informed consent from the subjects in accordance with the Declaration of Helsinki.

5.1 | Setup

Twelve surgeons (2 females, 10 males, and aged 30 to 62 years old) participated in the first set of experiments. The experimental set (Figure 2) comprised a virtual environment interface, a Sigma haptic device (Force Dimension, Switzerland) and a Phantom Omni haptic device (Geomagic, USA). The haptic control loop was implemented as a Robotic Operating System (ROS) application in C++ on a Linux computer. The Virtual Robot Experimentation Platform (VREP) was used for the development of the virtual environment. The haptic control rate was 1 kHz and the display rate was approximately 50 Hz. The simulated task was a simplification of surgical tissue-cutting using a unipolar electrocauter. The subjects were asked to cut along a reference curved path (the green curve in Figure 2) using the right haptic device (all subjects were right-handed). The subjects were asked to keep the left tool close to the cutting region and use it as a tensioning instrument. Both tools reflected interaction forces with the environment but only the cutting tool was subject to guidance AC. The tissue had a periodic 0.3 Hz translation that replicated respiratory motion. Touching the tissue with the cutting tool left a red mark on it. The simulated objects were non-deformable. All subjects were given as much time as they needed to train with the setup before the acquisition started (28 minutes on average). Each subject performed the task twice, with 3 assistive methods and a no-assistance case, resulting in a total of 8 repetitions per subject. The order of the acquisitions was random.

The second set of experiments were designed to extract the force characteristic of the guidance AC methods. Since these methods are prone to display undesired behaviour on boundary crossing, a straight line was used as the AC's geometry because it tended to increase the number of crossings in the preliminary tests. To narrow down the variability in motion of the tool, a sphere moving from one end of the line to the other with a constant velocity was displayed and the user was instructed to move the tool (a semi-transparent sphere) along the line accurately while following the sphere. Six non-expert users participated in this set of experiments.

5.2 | AC methods

The guidance active constraints used were dynamic non-energy-storing methods introduced by Kikuuwe et al.,[12] Bowyer and Baena[29] and Enayati et al.,[14] here referred to as plastic (P), plastic with redirection (PR) and viscous with redirection (VR) respectively. The key feature of these AC enforcement methods is that they engage only when the operator moves the tool. Thus, unlike energy-storing methods such as elastic AC enforcement, these methods are inherently unable to produce motion when the subject relaxes/releases the grip. AC methods with redirection guide the user towards the closest point of the reference by attempting to redirect the motion of the tool tip, while the P method only opposes orthogonal deviation. Descriptions of these methods are available,[14] and each method was implemented and its parameters were chosen as described in the corresponding paper. The parameters of the implemented ACs are reported in Table 1. Due to the innate difference of the elastic AC from the non-energy-storing ACs, it was not used in the tissue-cutting tasks. However, in the second experiment that was designed for analysis of force quality, the unconstrained trial was replaced by a simple elastic AC that generated a force towards the closest point of the path with a magnitude proportional to the distance of the tool tip from that point. The reason is that, although elastic AC may undermine safety due to motions generated by stored energy,[12] it exhibits favorable force behaviours such as minimum bounces on boundary crossings and intuitive proportionality. Therefore, if implemented with a reasonable elastic coefficient and a small amount

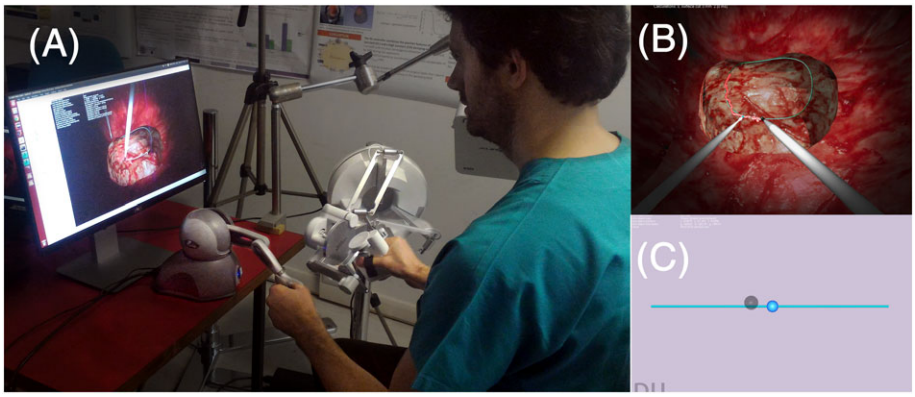

FIGURE 2 A, the experimental setup. B, the first virtual task was a simulation of surgical tissue cauterization along a 3D path on a periodically moving tissue. C, in the second task the user had to move along a straight line with a semi-transparent sphere while matching the constant velocity of a target sphere

TABLE 1 Parameters of the implemented AC methods

Kikuuwe et al.		Bowyer et al.		Enayati et al	
Param.	Value	Param.	Value	Param.	Value
R	5 N	f_c	5 N	F	5 N
F	0.5 N	σ_0	1667 N m^{-1}	B	80 N s m^{-1}
K	1667 N m^{-1}	σ_1	0 N s m^{-1}	D_m	0.002 m
B	2.5 N s m^{-1}	σ_2	2.5 N s m^{-1}		

of damping, it can be considered as a reference in analysis of force quality around the constraint boundary. The parameters of each guidance AC used in the experiments were selected so as to have a maximum force of 5 N. Although setting higher values could improve the enhancement in accuracy, it was decided to design the ACs to be more transparent and suggestive rather than restrictive. To prevent abrupt engagement of the constraints in boundary crossings, the damping and elastic coefficients of the VR and PR methods were linearly decreased to zero as the tool-tip approached the path from a distance of 2 mm until the crossing, therefore creating a gradual cylindrical dead-zone.

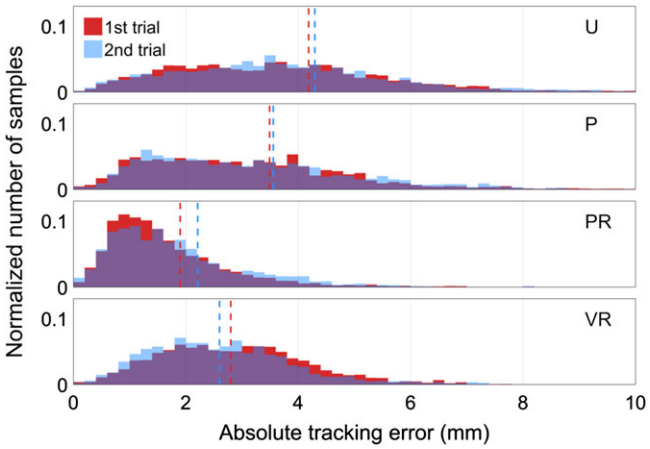

FIGURE 3 Population distribution of tracking error for each method and trial of the tissue-cutting task. The RMSE value of each trial is shown with a dashed line. A general accuracy enhancement can be observed for the acquisitions under guidance AC, especially for the PR and VR methods. While a small deterioration and improvement can be seen respectively for the PR and VR methods from the first trial to the second, no statistically significant difference is found between the trials of each method

5.3 | Results and discussion

In the following, we will employ the metrics described in this paper to analyze the results of the experiments. The results of the tissue-cutting experiment will be studied in terms of accuracy and subjective aspects, and the results of the line-tracking experiment will be investigated for evaluation of force quality.

5.3.1 | Accuracy analysis

Figure 3 depicts the distribution of the tracking error of the tool tip position for both repetitions of the tissue-cutting task. Each distribution represents the total population of the absolute tracking error samples of the 12 subjects performing a task with an AC method. It can be seen that the error distribution does not vary significantly between the first and second trials for the unconstrained and plastic cases. For the VR method a small shift is observed towards lower error values, resulting in an RMSE improvement from 3.1 mm in trial one to 2.9 mm in trial two. This can be related to users learning and acquiring more skills between the first trial and the second. Learning can be viewed from two aspects: acquiring more skills in interacting with the active constraint or becoming more familiar with the virtual environment and the hardware setup. If a learning effect of the latter type is present, the improvement is expected to be visible for most of the methods, which is not the case. In fact, the RMSE of the PR method increases from 2.0 mm for trial one to 2.2 mm for trial two. It can be hypothesized that since constraining the velocity is less intuitive than constraining the position, the subjects became abler in benefiting from the VR method in the second trial. Nevertheless, for none of the methods is a statistically significant difference found ($p < 0.05$) between the subjects' tracking RMS errors of the first and second trials. Therefore, to increase the sample size the two repetitions are

merged for the rest of the analysis. The Friedman test was used to study the statistical significance of the distributions. The test is a non-parametric version of two-way ANOVA that compares the populations, adjusting for possible subject effects.

Table 2 reports the RMSE, IQR and maximum error value for the tissue-cutting task. The values have been calculated from the concatenation of all the acquired tracking error signals for each method, i.e. 24 acquisitions (12 users and 2 repetitions per user) per method. As expected, the active constraints improve the accuracy compared to the unconstrained acquisitions. Among the ACs, the PR method shows a better enhancement in accuracy in terms of RMSE, followed by the VR method. A similar trend is present for the maximum absolute error. The IQR values show a better reduction in the error variation for the methods with motion redirection (PR and VR).

More information can be extracted from the acquired data by analyzing the variability in performance among subjects. Figure 4 depicts the distribution of the RMSE and maximum error for subjects. Each box plot represents the distribution of 24 values (RMSE or maximum value yielded from an acquisition) per method. The RMSE distribution reveals that while the IQR has decreased when the methods with redirection have been employed (0.7 mm for the unconstrained case (U) and 0.6 mm and 0.5 mm for the PR and VR methods respectively) the reduction is not substantial. This can be explained by the nonbinding nature of these guidance constraints and the relatively low force they generate. In other words, since the implemented ACs do not aim to strictly limit the tool's motion and rather transparently encourage it towards the desired reference, the subjects are still able to counteract the guidance force, and therefore the variability that partly stems from their skills and abilities is not considerably reduced. Nevertheless, regarding the maximum error, a considerable decrease is seen in the interquartile range from the unconstrained acquisitions to the PR and VR methods (3.5 mm for U, 1.6 mm for PR and 1.4 mm for VR). The results show that the PR and VR methods, while non-significant among themselves, are both significantly different ($p < 0.05$) from the U and P cases. Although an enhancement can be seen in terms of both RMSE and maximum error for the P method compared to the unconstrained acquisitions, the difference does not satisfy the statistical requirement. It can be argued that, for dynamic environments, redirecting the motion of the tool plays an important role, because that is the main difference between the P and the PR and VR methods.

TABLE 2 Accuracy enhancement metrics for the tissue-cutting task

AC methods	U	P	PR	VR
RMSE (mm)	4.2	3.5	2.1	2.7
IQR (mm)	2.6	2.4	1.3	1.6
Max. (mm)	14.4[a]	11.6	6.9[b]	7.8

[a]Ignoring an outlier at 18.3 mm.

[b]Ignoring an outlier at 14.8 mm.

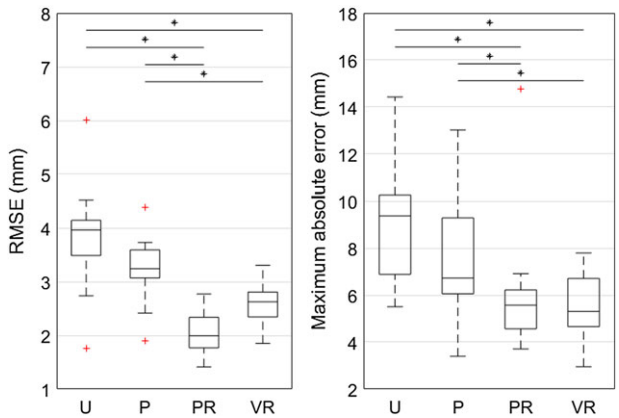

FIGURE 4 The RMSE and maximum error distributions among acquisitions of subjects. Each box plot contains 24 acquisitions per method. Central lines are medians; box edges are the 25th and 75th percentiles; whiskers are values within 1.5 times the IQR, plus markers are outliers and star-marked horizontal lines depict statistical significance

5.3.2 | Subjective analysis

At the end of each acquisition of the tissue-cutting task the surgeons were asked to express the distraction they felt from the method used in that acquisition by selecting a number from 0 to 10, respectively corresponding to no distraction at all and unbearably distracting. By definition, the unconstrained acquisitions were set to zero distraction. We tried to make it clear that the performance of the participant in an acquisition should not affect their judgement on distraction and that their response must be only regarding the forces applied to their hand from the guidance ACs. The average values of the distraction evaluation, reported in Table 3, show that the PR method received the worst distraction evaluation from the subjects. As some of the subjects expressed, the low evaluation was mainly due to jumps in the force magnitude on crossings of the reference path, which often caused small oscillations in the motion. This basic subjective evaluation of the force will be confronted with the objective analysis of the generated forces in the next subsection. Fleiss' kappa was used to measure inter-rater reliability. The value was calculated as 0.44, which according to the interpretation guidelines of Landis and Koch[26] indicates a moderate agreement among the subjects. The average execution times reported in Table 4 depict a considerable reduction when guidance ACs were used, especially with the PR and VR methods. The faster execution time corresponding to these methods can suggest their effectiveness in reducing the cognitive load on the subjects.

TABLE 3 Evaluation of the subjects regarding the distraction level of each method

AC methods	U	P	PR	VR
Distraction level	0[a]	3.5	6.3	3.0

[a]The unconstrained case was assigned zero distraction by default.

TABLE 4 Average task execution time

AC methods	U	P	PR	VR
Time (s)	54.0	39.9	34.3	33.5

5.3.3 | Force analysis

Although the maximum force magnitude was equal (5 N) for all the three ACs, the plastic methods reached the maximum far more often than the viscous AC. This can be seen in the average forces administered by the ACs in the tissue-cutting task, which are 1.7 N, 2.6 N and 1.5 N respectively for the P, PR and VR methods. The analysis of the second experiment's results can yield a better understanding of the force characteristics of the AC methods. An elastic (E) AC was used so as to have a reference for comparison. Despite the risks related to the stored energy in elastic constraint, the forces administered by such a constraint exhibit a desirable proportionality in magnitude and smooth variation in direction. As explained in section III, the derivatives of force (as a measure of jerk), and the derivatives of azimuth and polar angles represent the variability of the generated force, and therefore are desired to be limited for guidance ACs. Figure 5 shows a 3 s time series of these values for each AC method, acquired during one of the trials of the second experiment. It is evident that while the P and VR methods exhibit behaviours similar to that of the elastic method, the PR method possesses large magnitudes and variability of the derivatives of both force and angles. As mentioned before, method P does not act against motions parallel to the reference path, which makes it less enforcing in a task of following a straight line. In contrast, the relatively higher magnitude of the polar rate of the VR method shows greater engagement of this method, which is mainly because of motion redirection on the reference path boundary. The averages of the absolute values of force and angle derivatives for all acquisitions reported in Table 5 confirm what can be observed from Figure 5. The PR method exhibits large variations in force magnitude and direction that can lead to oscillations and distracting artifacts. It is not surprising, however, to observe high rates of force magnitude for the plasticity-based methods because they are designed to simulate plastic deformation that ideally approaches a step function of the tool displacement. The same conclusions can be drawn from the frequency domain analysis of the force signals depicted in Figure 6. While the power of elastic, P and VR methods decreases monotonically as the frequency increases, the PR method's power shows a few local peaks that are due to boundary oscillations generated in the line-following task of the second experiment. The higher energy of the force magnitude of plastic methods is explained by their higher average force magnitude as discussed earlier in this subsection. Note that the second experiment serves as a testbed for assessing the quality of the guidance force in a worst-case scenario, and the undesired behaviours would be less intense for most practical applications. Nonetheless, these results are in agreement with the subjective evaluation of the distraction levels. As can be seen from the results, there is necessarily no connection between force quality and accuracy. Although the subjects achieved

TABLE 5 Time domain metrics in force analysis

AC methods	E	P	PR	VR
Mean absolute force derivative (N s^{-1})	1.9	2.7	31.8	2.8
Mean absolute azimuth rate (rad s^{-1})	0.4	0.9	8.4	0.8
Mean absolute polar rate (rad s^{-1})	1.3	0.7	17.1	2.4

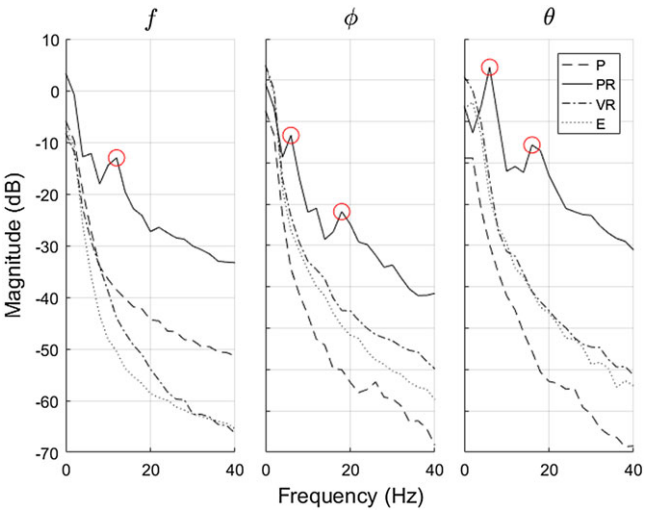

FIGURE 6 Estimate of Welch power spectral density of the force magnitude and its spherical angles. The circled peaks denote the presence of oscillations in forces generated by the PR method in the line-following task

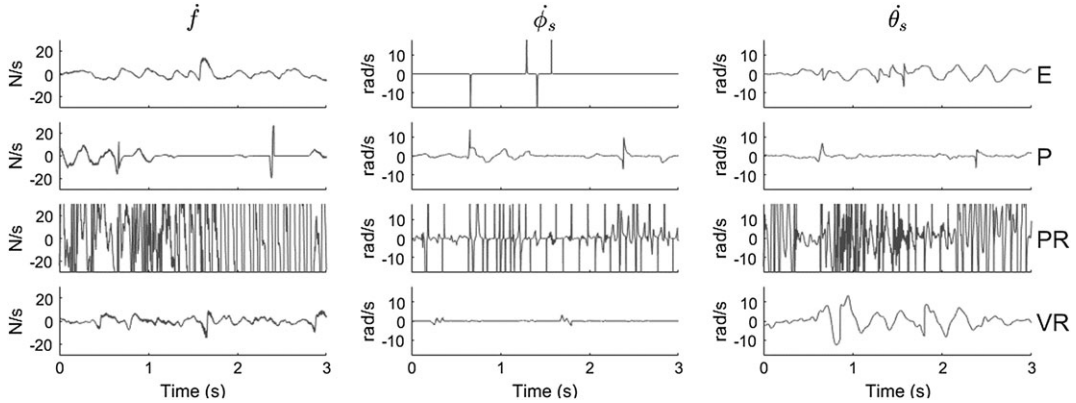

FIGURE 5 An example plot of 3 s time-domain metrics from the second experiment's acquisitions for qualitative assessment. The vertical axes are cropped for clarity. While the P and VR methods show moderate rates comparable to those of the elastic AC, the PR method exhibits high magnitudes and frequencies of force and angular rate. The sections with zero $\dot{f}$ that can be seen in the P method's graph are due to the inactivity of this method in parallel tool motions

higher accuracy in the tracking task with the PR method than with the other methods, they reported high distraction levels for that method, which is confirmed by the force analysis. This in turn demonstrates the necessity of in-depth and more inclusive analysis of methods of force generation in guidance AC.

It must be noted that the final evaluation of these methods needs to involve application-specific aspects too. For example, in the experiments performed in this work, the viscosity-based AC demonstrated good accuracy enhancement relative to the other implemented methods, but for a task involving low velocities, such an enforcement method may not result in satisfactory improvements in accuracy and may even lead to safety concerns.

6 | CONCLUSIONS

Success in robotic surgery is largely dependent upon precise and delicate manipulation of tissues. The performance in such tasks can be augmented by implementing guidance active constraints not only to avoid tissue damage but also to reduce the cognitive load on the surgeons. These methods should be intuitive to use and must not interfere with the clinicians' intended actions. Although various methods of guidance AC enforcement have been introduced in the literature, the common metrics for evaluating these methods are far from exhaustive. In this work, we introduced a set of metrics that can be employed to investigate the principal parameters and characteristics of guidance ACs. Furthermore, the application of these metrics and the importance of an all-encompassing evaluation was depicted through two sets of experiments. Since these methods are designed to be used by humans, it is critical to consider task-related subjective aspects and overall human factors. It was observed through subjective and force analyses that an enforcement method might help subjects attain higher tracking accuracies, while it showed higher levels of disturbance. Although achieving high accuracy despite high disturbance may seem counter-intuitive, it must be considered that an enforcement method can encourage the tool to follow the reference path and reduce the tracking error, yet can do so through the application of rapidly varying and distracting forces that do not leave a trace in accuracy analysis. Such disturbance may not introduce explicit safety risks in some applications, but they can still agitate and disturb the operator and affect the user experience. As an overall conclusion, it is encouraged to include a force/torque analysis (even simply measures of the derivative of the forces/torques generated by an AC) along with a subjective assessment directed at the enforcement quality, when assessing the performance of novel AC methods.

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Horizon 2020 Industrial Leadership program under grant agreement No. H2020-ICT-26-2016-1-732515 (SMARTsurg project). No conflict of interest is reported.

ORCID

Nima Enayati http://orcid.org/0000-0001-5337-9446

REFERENCES

1. Shademan A, Decker RS, Opfermann JD, Leonard S, Krieger A, Kim PCW. Supervised autonomous robotic soft tissue surgery - supplementary material. *Sci Transl Med*. 2016;8(337). https://doi.org/10.1126/scitranslmed.aad9398
2. Murali A, Sen S, Kehoe B, et al. Learning by observation for surgical subtasks: multilateral cutting of 3D viscoelastic and 2D orthotropic tissue phantoms. *Int Conf Robot Autom*. 2015;1202-1209. https://doi.org/10.1109/ICRA.2015.7139344
3. Payne CJ, Yang G-Z. Hand-held medical robots. *Ann Biomed Eng*. 2014;42(8):1594-1605. https://doi.org/10.1007/s10439-014-1042-4
4. Mofidi A, Plate JF, Lu B, et al. Assessment of accuracy of robotically assisted unicompartmental arthroplasty. *Knee Surg Sports Traumatol Arthrosc*. 2014;22(8):1918-1925. https://doi.org/10.1007/s00167-014-2969-6
5. Penza V, De Momi E, Enayati N, Chupin T, Ortiz J, Mattos LS. EnViSoRS: Enhanced vision system for robotic surgery. A user-defined safety volume tracking to minimize the risk of intraoperative bleeding. *Front Robot AI*. 2017;4:1-15. https://doi:10.3389/frobt.2017.00015
6. Enayati N, De Momi E, Ferrigno G. Haptics in robot-assisted surgery: Challenges and benefits. *IEEE Rev Biomed Eng*. 2016;9:49-65. https://doi.org/10.1109/RBME.2016.2538080
7. Bowyer SA, Davies BL, Rodriguez Y, Baena F. Active constraints/virtual fixtures: a survey. *IEEE Trans Robot*. 2014;30(1):138-157. https://doi.org/10.1109/TRO.2013.2283410
8. Leibrandt K, Marcus HJ, Kwok K-W, Yang G-Z. Implicit active constraints for a compliant surgical manipulator. *2014 IEEE Int Conf Robot Autom*. 2014;276-283. https://doi.org/10.1109/ICRA.2014.6906622
9. Zhang D, Zhu Q, Xiong J, Wang L. Dynamic virtual fixture on the Euclidean group for admittance-type manipulator in deforming environments. *Biomed Eng Online*. 2014;13(1):51. https://doi.org/10.1186/1475-925X-13-51
10. Ryd F, Chizeck HJ. Forbidden-region virtual fixtures from streaming point clouds: remotely touching and protecting a beating heart. *IEEE Int Conf Intell Robots Syst*. 2012:3308-3313.
11. Mylonas GP, Kwok KW, James DRC, et al. Gaze-contingent motor channelling, haptic constraints and associated cognitive demand for robotic MIS. *Med Image Anal*. 2012;16(3):612-631. https://doi.org/10.1016/j.media.2010.07.007
12. Kikuuwe R, Takesue N, Fujimoto H. A control framework to generate nonenergy-storing virtual fixtures: use of simulated plasticity. *IEEE Trans Robot*. 2008;24(4):781-793. https://doi.org/10.1109/TRO.2008.924947
13. Bowyer SA, Rodriguez Y, Baena F. Dynamic frictional constraints for robot assisted surgery. *World Haptics Conf*. 2013;319-324. https://doi.org/10.1109/WHC.2013.6548428
14. Enayati N, Costa ECA, Ferrigno G, De Momi E. A dynamic non-energy-storing guidance constraint with motion redirection for robot-assisted surgery. In: *IEEE International Conference on Intelligent Robots and Systems*. Vol 2016-November. Daejeon, Korea; 2016:4311–4316. doi:https://doi.org/10.1109/IROS.2016.7759634.
15. Navkar NV, Deng Z, Shah DJ, Bekris KE, Tsekos NV. Visual and force-feedback guidance for robot-assisted interventions in the beating heart with real-time MRI. *2012 IEEE Int Conf Robot Autom*. 2012;689-694. https://doi.org/10.1109/ICRA.2012.6224582
16. Gibo TL, Verner LN, Yuh DD, Okamura AM. Design considerations and human-machine performance of moving virtual fixtures. *2009 IEEE Int Conf Robot Autom*. 2009;671-676. https://doi.org/10.1109/ROBOT.2009.5152648
17. Takesue N, Kikuuwe R, Sano A, Mochiyama H, Fujimoto H. Tracking assist system using virtual friction field. *2005 IEEE/RSJ Int Conf Intell Robot Syst*. 2005;3927-3932. https://doi.org/10.1109/IROS.2005.1545292

18. Bettini A, Marayong P, Lang S, Okamura AM, Hager GD. Vision-assisted control for manipulation using virtual fixtures. *IEEE Trans Robot*. 2004;20(6):953-966. https://doi.org/10.1109/TRO.2004.829483
19. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. *Geosci Model Dev*. 2014;7(3):1247-1250. https://doi.org/10.5194/gmd-7-1247-2014
20. Payne CJ, Marcus HJ, Yang G-Z. A smart haptic hand-held device for neurosurgical microdissection. *Ann Biomed Eng*. 2015;43(9):2185-2195. https://doi.org/10.1007/s10439-015-1258-y
21. Flash T, Hogan N. The coordination of arm movements: an experimentally confirmed mathematical model. *J Neurosci*. 1985;5(7):1688-1703. http://www.ncbi.nlm.nih.gov/pubmed/4020415. Accessed March 30, 2016
22. Marchal-Crespo L, Reinkensmeyer DJ. Review of control strategies for robotic movement training after neurologic injury. *J Neuroeng Rehabil*. 2009;6(1):20. https://doi.org/10.1186/1743-0003-6-20
23. Schloerb DW. A quantitative measure of telepresence. *Presence Teleoperators Virtual Environ*. 1995;4(1):64-80. http://web.mit.edu/schloerb/www/publications/schloerb-6.pdf
24. Koehn JK, Kuchenbecker KJ. Surgeons and non-surgeons prefer haptic feedback of instrument vibrations during robotic surgery. *Surg Endosc*. 2015;1-14. https://doi.org/10.1007/s00464-014-4030-8
25. Slater M. How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence Teleoperators Virtual Environ*. 2004;13(4):484-493. https://doi.org/10.1162/1054746041944849
26. Gwet K. *Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement between Two or Multiple Raters*. STATAXIS: Gaithersburg, MD; 2001.
27. Hart SG, Staveland LE. Development of NASA-TLX (task load index): results of empirical and theoretical research. *Adv Psychol*. 1988;52:139-183. https://doi.org/10.1016/S0166-4115(08)62386-9
28. Hart SG. NASA-task load index (NASA-TLX); 20 years later. *Proc Hum Factors Ergon Soc Annu Meet*. 2006;50(9):904-908. https://doi.org/10.1177/154193120605000909
29. Bowyer SA, Baena FRY. Dissipative control for physical human-robot interaction. *IEEE Trans Robot*. 2015;31(6):1281-1293. https://doi.org/10.1109/TRO.2015.2477956

How to cite this article: Enayati N, Ferrigno G, De Momi E. Performance metrics for guidance active constraints in surgical robotics. *Int J Med Robotics Comput Assist Surg*. 2018;14:e1873. https://doi.org/10.1002/rcs.1873