Journal of Robotic Surgery (2024) 18:113
https://doi.org/10.1007/s11701-024-01868-z

RESEARCH

Evaluation of different approaches to define expert benchmark scores for new robotic training simulators based on the Medtronic HUGO™ RAS surgical robot experience

Mark Brentnall[1] · John Lenihan Jr.[2] · Christopher Simmonds[3] · Anand Malpani[1] · Antonio R. Gargiulo[4] · Martin Martino[5] · Jeffrey S. Levy[6]

Received: 9 January 2024 / Accepted: 10 February 2024

Abstract

New robot-assisted surgery platforms being developed will be required to have proficiency-based simulation training available. Scoring methodologies and performance feedback for trainees are currently not consistent across all robotic simulator platforms. Also, there are virtually no prior publications on how VR simulation passing benchmarks have been established. This paper compares methods evaluated to determine the proficiency-based scoring thresholds (a.k.a. benchmarks) for the new Medtronic Hugo™ RAS robotic simulator. Nine experienced robotic surgeons from multiple disciplines performed the 49 skills exercises 5 times each. The data were analyzed in 3 different ways: (1) include all data collected, (2) exclude first sessions, (3) exclude outliers. Eliminating the first session discounts becoming familiar with the exercise. Discounting outliers allows removal of potentially erroneous data that may be due to technical issues, unexpected distractions, etc. Outliers were identified using a common statistical technique involving the interquartile range of the data. Using each method above, mean and standard deviations were calculated, and the benchmark was set at a value of 1 standard deviation above the mean. In comparison to including all the data, when outliers are excluded, fewer data points are removed than just excluding first sessions, and the metric benchmarks are made more difficult by an average of 11%. When first sessions are excluded, the metric benchmarks are made easier by an average of about 2%. In comparison with benchmarks calculated using all data points, excluding outliers resulted in the biggest change making the benchmarks more challenging. We determined that this method provided the best representation of the data. These benchmarks should be validated with future clinical training studies.

Keywords Surgical education · Robotic surgery · Virtual simulation · Proficiency-based training · Benchmarking simulation passing scores

✉ John Lenihan Jr.
jlenihan@comcast.net; Dr.John.Lenihan@gmail.com

Mark Brentnall
brentnallm@gmail.com

Christopher Simmonds
ccrsimmonds@gmail.com

Anand Malpani
Malpani.anand.89@gmail.com

Antonio R. Gargiulo
agargiulo@bwh.harvard.edu

Martin Martino
medmartino@gmail.com

Jeffrey S. Levy
jeffslevy@gmail.com

[1] Surgical Science, Seattle, WA, USA

[2] University of Washington School of Medicine, Seattle, WA, USA

[3] New Canaan, CT, USA

[4] Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA

[5] Ascension Medical Group, Jacksonville, FL, USA

[6] Temple University, Philadelphia, PA, USA

Published online: 07 March 2024

Introduction

Virtual reality simulators for robotic surgery are widely accepted as valuable tools for training and skills maintenance [1, 2]. New robot-assisted surgery platforms being developed will be required by societies and regulatory agencies to have proficiency-based simulation training available either as an embedded feature or as a stand-alone product [3]. As new simulation exercises are developed for each new robotic platform that is being introduced to the market, there is a need to identify valid methodologies for determining the appropriate passing thresholds for each performance metric. Unique benchmarks should be established for each new surgical robot. In the past, this was done gradually over several years as new surgeon users contributed to the scoring data. However, scoring systems and performance feedback for trainees are not consistent across all simulator platforms. Prior publications describing how commercial simulator benchmark scores have been derived are lacking. A few papers outlining how single centers used local expert surgeons to establish passing benchmarks [4]. Typically, these methods defined passing as 1 standard deviation above the expert's mean scores.

The concept of "Proficiency Based Training" has been promoted as the best way to ensure that a new surgeon has achieved a high level of performance using a robotic surgical platform [5]. Dreyfus and Dreyfus in 1986 described how a student typically progresses from being a novice, to becoming eventually "competent" (safe), and ultimately "proficient" (good) at any required skill, such as operating a surgical robot [6]. This classic model of skills acquisition has been utilized in all levels of medical training [7]. To ensure safe and efficient users, many robotic surgery training programs are now requiring that students demonstrate proficiency on a simulator utilizing deliberate practice and passing all the exercises in a specified curriculum multiple times before being allowed to transition to live patients in the operating room, or even expensive wet lab training using animal models or cadavers [8].

When looking at how a student demonstrates proficiency in a specific simulation exercise, there are three elements to consider:

First, one must determine the passing level or benchmark of each metric in the exercise that they are being asked to master. A typical simulation exercise may have between five and ten individual metrics, and these may be classified as either efficiency metrics (measures related to efficiency use of the system such as time to complete, distance traveled by instruments) or safety metrics (measures related to errors that could impact patient safety, such as misapplied energy, drops, instrument collisions, instruments out of view, etc.) The benchmark values for these metrics are typically pre-loaded into the simulator and are usually based on evaluation of expert performance or some data-driven algorithm.

Second, one needs to decide how metric scores are aggregated and considered to determine if the trainee has achieved a passing score for that exercise. Here again there are two common approaches.

One approach, which we term the "Classic Approach," is to produce a standardized value for each metric and then sum or average them, giving a total score that is a sum or average of all the performance metrics. A passing level is then set based on the values of all metrics that are included in that exercise. On some simulators, this is displayed as a percentage of some arbitrary maximum, with 100% being the highest achievable and 80% often considered a passing score. This method does not acknowledge the performance of each metric in that exercise, and it is not always clear how this 80% passing threshold is determined. Another problem with this approach is that when all metrics are averaged into a combined score, that method allows trainees to compensate for poor performance in some metrics by doing exceptionally well in others. For example, going very fast and doing well in the time and efficiency metrics may compensate for making lots of errors and doing poorly in the safety metrics while still obtaining an over-all average score that is considered "passing".

A different approach is to force the trainee to pass every single "critical" metric before they can pass each exercise. If the student fails just one metric, either an efficiency or safety metric, they will not pass that individual exercise. This can, at times, be frustrating for the trainee, who may be doing well in most areas but failing in just one or two critical metrics. However, this method ensures that all trainees are required to strive for excellence in both efficiency as well as safety metrics. We have referred to this as the "Absolute" method. The new Hugo™ RAS Simulator uses an "Absolute" approach for assessing pass/fail of each exercise. This method has been utilized and validated in numerous studies on previous simulators [9, 10]. Historically, the first wisely used commercial robotic simulator, the Mimic Technologies (Seattle, WA) dV-Trainer™ utilized the Absolute Scoring method. The Intuitive Backpack Simulator was later developed and initially utilized the same software and scoring methodologies. Later, however, Intuitive Surgical (Sunnyvale CA) developed their own simulator called the Intuitive Skills Simulator™ which used many of the legacy Mimic technologies exercises and several of their own internally developed exercises, but moved to the Classic Scoring methodology for all exercises. The Symbionix (Tel Aviv, Israel)

Robotix-Mentor Simulator currently uses a hybrid scoring system utilizing both Classic and Absolute methodologies.

Third, you need to determine how many passes of an exercise constitute “proficiency.” Requiring multiple passed attempts promotes the idea that a proficient user should consistently achieve the performance standards of the metric benchmarks, and not “just get lucky once”. Some academic simulation centers today only require that a student passes each exercise once. We advocate for passing each exercise multiple times to ensure that the student has mastered the exercise and achieved the desired level of proficiency. However, depending on the complexity of the exercise, and the difficulty of passing the metrics within the exercise, this should be customizable by each institution based on the training goals and the population being trained.

Once the definition of “proficiency” is determined, based on the three criteria listed above (metric benchmarks, critical metrics, and required number of “passing” attempts), the next task is to calculate those passing score benchmarks on the simulator, ideally using expert surgeon performances. There are many things that need to be taken into consideration for this process. Selecting the appropriate data for benchmarking analyses can included the following methods: (1) include all collected metric data, (2) exclude data from initial attempts at each exercise, or (3) exclude data identified as outliers. Outlier scores typically have nothing to do with the expert surgeon’s performance, but often are related to the testing environment, such as computer failures, software glitches, phone calls, and other distractions. These interruptions can affect the outcomes of the data analysis if those sessions are included. In our analysis, we looked at each of these three methods to determine which approach would provide the best data to determine the best way to determine correct expert benchmark scores.

Methods

Nine surgeons from multiple disciplines who were each very experienced in robotic surgery on a similar existing platform participated in this study designed to collect data and create the benchmark passing scores for the skills training exercises on the new Hugo™ RAS Simulator. After a short familiarization process, each of the 9 surgeons was asked to perform five sessions (exercise attempts) for each of the 49 available simulation exercises covering both basic and advanced skills. The 49 exercises were divided into two curricula: BM1 and BM2, which were completed over two consecutive days, one day for BM1 and one day for BM2. These exercises were focused on basic robotic surgery skills (BM1), including camera movement, clutching, use of energy; and advanced skills (BM2), such as sewing, knot tying, dissection, and the 6 FRS (Fundamentals of Robotic Surgery) skills exercises. Of the 9 surgeons, 7 completed the entire protocol, BM1 and BM2. One surgeon only completed BM1, and another only completed BM2; so we included these results as a single complete surgeon data set for the benchmarking analysis. Each individual simulation exercise has from 6 to 10 performance metrics, depending on the relevance to the task, which were categorized either as safety metrics or efficiency metrics. All individual metrics were included in the analyses. With 8 complete surgeon data sets, 5 sessions for each exercise, and 49 total exercises, we expected 1960 sessions and 40 data points for each metric in each exercise. The actual total number of sessions collected was 1915. A few exercise-metrics only had 38 or 39 data points as some surgeons inadvertently skipped or did not complete all 5 sessions for one or two exercises during their two-day benchmark collection. We concluded that this did not dramatically affect the results of the data analysis, and the collected data was sufficient to create the benchmarks.

The data was initially collected on the simulator in SQLite database, then exported for aggregation and analysis using Tableau™ (Tableau Software, Seattle, WA). Subsequent analysis on the benchmarks calculated in Tableau™ was done using Excel™ (Microsoft Corporation, Redmond, WA). This data is available to the public and can be exported directly for off-line analysis or references through the user interface on all Hugo™ RAS Simulators since 2022.

As previously mentioned, the data were analyzed in 3 different ways: (1) include all data collected, (2) exclude first sessions of each exercise per surgeon, and (3) exclude only outlier values. Outliers were identified using a common statistical technique involving the interquartile range of the data using Box and Whisker plots [11]. (Fig. 1.) From the remaining data using each method above, MEAN and standard deviations (SD) were calculated, and the passing benchmark was set at a value of 1 SD above the MEAN. It should be noted that, for every metric in these exercises, larger values indicate worse performance (more time, greater instrument movements, more errors, etc.). Also, in every case of the outlier determination, the calculated outliers were a result of exceptionally worse performance. Our calculations did not identify any outliers at the front end of the data points.

We then did sub-analyses looking at potential outcomes in passing the training exercises if we used the three different methods of calculating benchmark scores.

Results

Nine multi-specialty expert robotic surgeons were able to perform 49 exercises each over a two-day period. This resulted in 1,915 valid exercise sessions (58 + hours of simulation time) and produced 13,279 individual metric scores

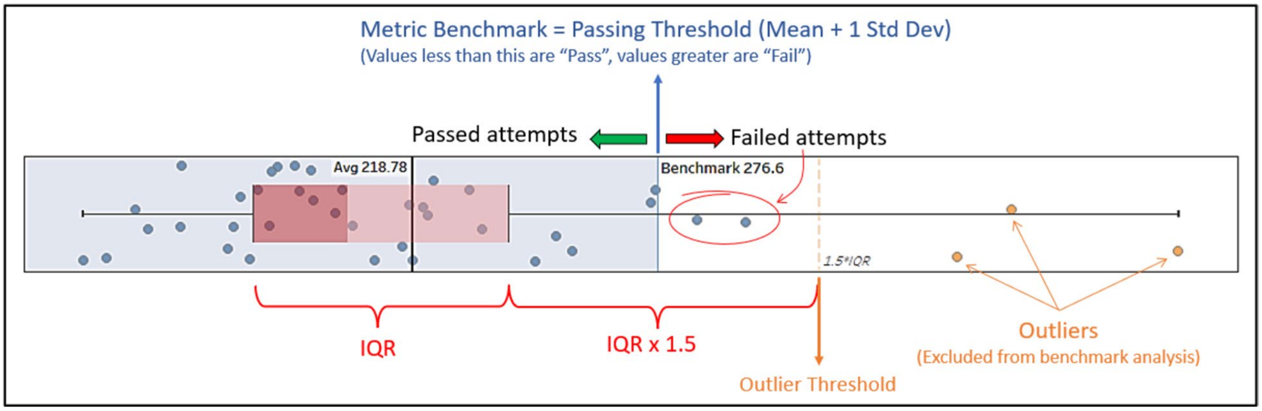
Metric Benchmark = Passing Threshold (Mean + 1 Std Dev)
(Values less than this are "Pass", values greater are "Fail")
Passed attempts
Failed attempts
Avg 218.78
Benchmark 276.6
1.5*IQR
IQR
IQR x 1.5
Outliers
(Excluded from benchmark analysis)
Outlier Threshold

Fig. 1 Illustration of outlier calculations and resulting benchmark (Pass/Fail Threshold) for the metric "Economy of Motion" in the exercise "Dots and Needles 1"

spread across 35 different metrics. Not all exercises have the same number of metrics. The average number of metrics for each exercise is between 6 and 7 per exercise. It took surgeons between 5.6 and 8.6 h (average time was 7.3 h) to complete the two benchmark curricula, BM1 and BM2 (Fig. 2.). This data was used to establish the passing benchmarks for each metric in each exercise, which was set at 1 SD above the MEAN value calculated from the remaining data points after excluding the outliers. The entire process, from surgeon recruitment to final implementation of the calculated benchmarks, took just under three months.

The 49 exercises can be divided into 7 exercise categories based on skill set: Instrument Manipulation, Camera Control, Clutching, Energy Control, Fourth Arm Control, Needle Control and Driving, Suturing, Knot-tying, Dissection, Stapling, and FRS. These can be further grouped into four broad categories: 1. Foot Control exercises (including camera and energy); 2. Basic Skills (exercises focused on general

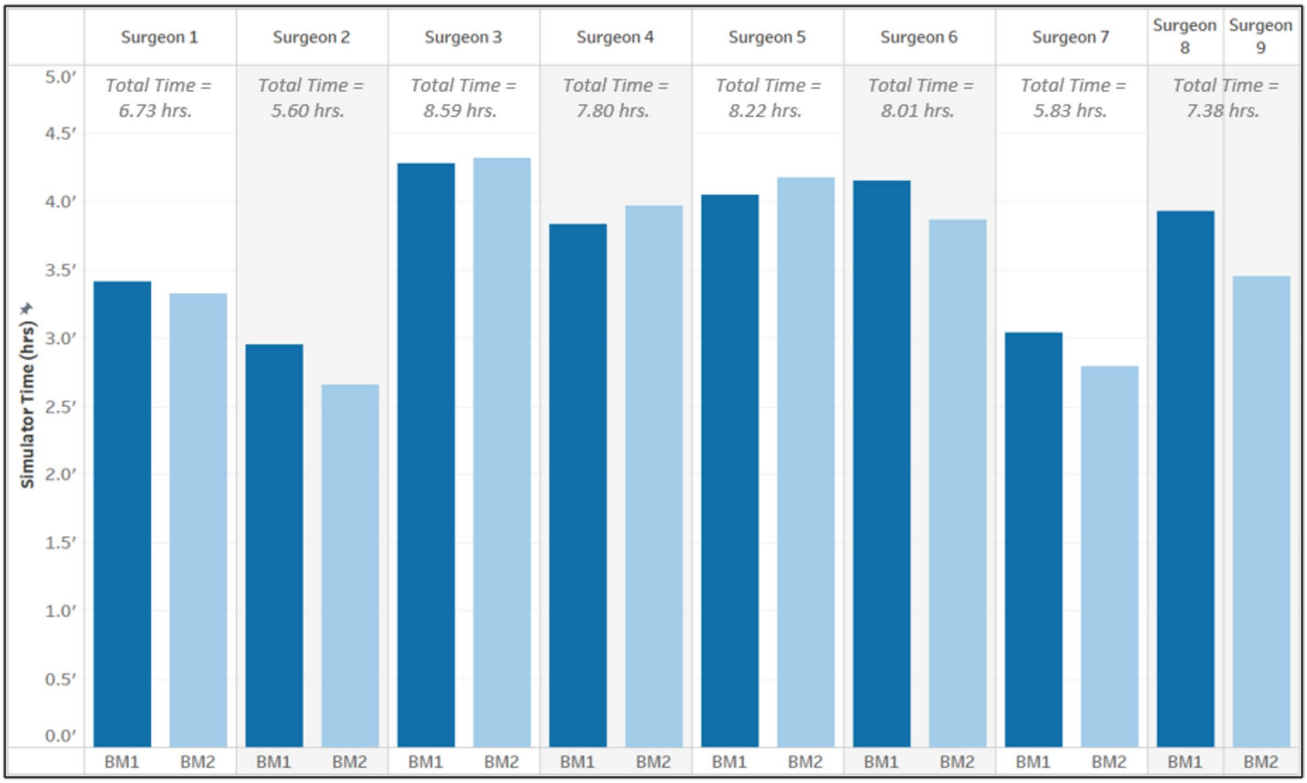

Fig. 2 Benchmark data collection summary: Time and Sessions Per Curriculum

instrument manipulations and concepts such as clutching); 3. Advanced Skills (where finer manual dexterity is required, including skills, such as needle-handling, suturing, and knot tying); and 4. Other (which includes the remaining exercises, focused on advanced instrumentation tasks such as applying clips and stapling as well as test protocols such as FRS.) Most of the time was spent on needle-handling and suturing exercises, likely due to the quantity of exercises, as well as the complexity and length of those skill tasks (Fig. 3).

Once the sessions were completed, the data were analyzed to help develop the benchmarks using the three methods previously stated. When all data were included (Method 1), nominally about 40 data points (5 sessions × 8 surgeons) were available for each metric for each exercise (a few exercises only had 38 or 39 points due to some surgeons missing some sessions). When every first session is excluded (Method 2), nominally only 32 data points are available (4 sessions × 8 surgeons). When outliers were removed (Method 3), the remaining data varied depending on the outlier calculations for those metrics, but on average, 38 data points are available. Method 3 removed fewer overall data points than Method 2, but it also had a greater impact on moving the benchmark thresholds. Using Method 3, the benchmarks were, on average, 11% more difficult to pass (smaller values for the pass/fail threshold) compared to the benchmarks from Method 1. Method 2 included 32 data points for each metric, and the benchmarks were approximately 2% easier to pass (larger pass/fail threshold values) compared to Method 1. (Table 1).

Method 2 (remove first sessions) assumes that the first attempts at a new exercise involve task familiarization, resulting in abnormally bad performance. The data did not support this assumption, since including the first sessions resulted in more difficult benchmarks, suggesting that the performance in those first sessions was already better than some of the subsequent performances. Method 3 (remove statistical outliers) assumes that there are external factors affecting some sessions (distractions, software glitches, etc.) and these are not representative of a "typical" attempt.

We then decided to do a further retrospective analysis and see if the surgeons would have passed their own establishing benchmarks during the sessions in which the data was

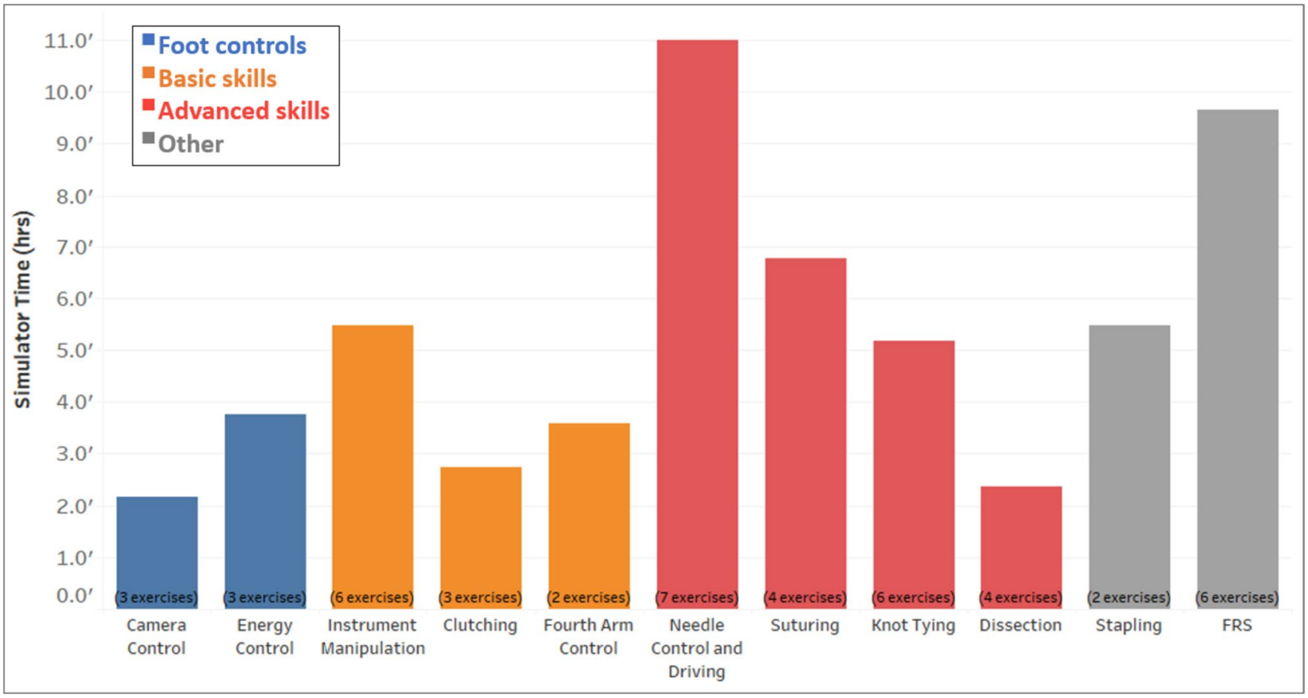

Fig. 3 Benchmark data summary: time per exercise category

Table 1 Available data points and relative passing rate when benchmarks are calculated 3 different ways

Analysis method:	Avg. number of data points per exercise metric	Avg. relative rate of passing
1. Include all data	40 (8 surgeon data sets × 5 sessions)	100%
2. Exclude first session	32 (8 surgeon data sets × 4 sessions)	102% (2% easier than Method 1)
3. Exclude outliers	38 (varies per exercise and metric)	89% (11% harder than Method 1)

collected. Interestingly, 15.53% of the metrics would have been failed by the expert surgeons. This is not completely surprising as these metrics are an average of scores produced by different people.

We tried to better understand what was driving the failures. We hypothesized it could be caused by three different factors. First, it could be the type of exercise that led to worse performance. Second, it could be the metrics themselves with some metrics being more difficult to pass than others. Lastly, it could be a variation between the surgeons, themselves. We will look at these separately.

Our first approach was to divide the exercises into the same groups that we had in Fig. 3. The hypothesis was that perhaps the advanced skill exercises would have more failures. We noted however that the percentage of failures within each group was relatively consistent. On average, 15.53% of the metrics were considered "Failed" according to the established benchmarks; however, there was little variation between the exercise groups, with a standard deviation of only 1.29%.

Next, we looked at the metrics themselves that were being measured. As previously discussed, these metrics were divided into safety and efficiency metrics. Safety metrics are related to activities that could potentially cause a patient harm while efficiency metrics are related to the speed and the smoothness of the surgeon's movements (Table 2.)

Although there are 35 different metrics used across all exercises, the 5 metrics in Table 2 are the ones used in most exercises and make up 78% of all the metrics captured. While there is a slightly larger variation between them than with the exercise groups, with Economy of Motion and Time to Complete showing the largest variation, the average failure rate is still relatively consistent, with a standard deviation between metrics of just 3.66%.

We also looked at the failure rate per surgeon where we noted a greater variation between surgeons and the number of metrics that were considered "failed" based on the calculated benchmarks. The average for all surgeons was 15.53%, but here, surprisingly, we had the largest standard variation between surgeons of 9.10%.

This was an unexpected finding since all the surgeons selected for the study have had significant experience using a similar, existing robotic system. We have seen variations like this in the past in when calculating a relative mean proficiency index for simulation users [12]. In this study, we saw significant variation in the time it took for surgeons to reach proficiency on a curriculum of 32 exercises. Studies have also shown that novice populations will score differently on a VR robotic surgery curriculum based on their innate ability [13]. It could be that these performance variations are tied to the innate ability of these experienced surgeons. This 15% fail rate is also a feature of a normal distribution curve, so were not totally unexpected [14]. As mentioned before, scoring metrics can be adjusted by each institution to allow for this effect and allow for appropriate pass rates for their users.

Finally, we estimated the impact of these differences on the likelihood of a surgeon passing or failing a curriculum. As we discussed above, there a currently two methodologies being used for scoring users once they have completed an exercise. The "CLASSIC" (or "Averaging") method uses the average of all scores and translates each metric into a percentage and then produces a total score by averaging those percentages. This is normally shown as a number with 100% being the highest achievable. We compared this to the "ABSOLUTE" method which defines passing only when the student passes each individual SAFETY and EFFICIENCY metric before they are allowed to pass an exercise. Even if they score highly on five of the six metrics, they will fail until they pass the final metric.

The example in Table 3 shows the difference between the two approaches for a single surgeon. In the case of this expert surgeon, if using the CLASSICAL approach, the user would pass the exercise with an average score of 85%; but if we used the ABSOLUTE approach where each metric is critical to passing, the user would fail since the user failed three individual metrics (Table 3.)

We then applied the two methodologies to the surgeons across all exercises. Using the AVERAGING methodology, 82% of the exercise would have been passed with the highest-scoring surgeon passing 95% of the exercises and lowest-scoring passing 55%. All surgeons would have passed each exercise at least once. Using the ABSOLUTE methodology, the pass rate drops to 48% with the highest-scoring surgeon passing 72% and the lowest-scoring passing only 20%. About 11% of the exercises would not have been passed once by the expert surgeons in the five attempts.

Table 2 The number of passed and failed metrics for all sessions per metric

Metric	Type	Passed	Failed	Grand total	Percentage "failed"
Economy of motion	Efficiency	1466	368	1834	20.07%
Time to complete exercise	Efficiency	1443	391	1834	21.32%
Master workspace range	Efficiency	1330	189	1519	12.44%
Instruments out of view	Safety	1525	309	1834	16.85%
Excessive instrument force	Safety	1391	203	1594	12.74%

Table 3 Comparison of Pass/Fail scores with CLASSIC & ABSOLUTE Methods of Scoring for One Surgeon's Attempt at one Exercise

Individual Metrics	ABSOLUTE Method	CLASSIC Method
Economy of Motion	Pass	92%
Excessive Instrument Force	Fail	70%
Instrument Collisions	Fail	75%
Instruments Out Of View	Pass	80%
Master Workspace Range	Pass	88%
Time to Complete Exercise	Pass	107%
OVERALL EXERCISE RESULT:	**Fail**	**85% (Pass)**

85% (Pass): This score meets the criteria of 1 standard deviation above the expert benchmark threshold scores for passing each exercise. With the classic method of averaging numerical scores for each metric, it is considered a pass even though the surgeon may have failed a critical safety metric. In the absolute method, however, each metric must be passed before the entire exercise can be considered "passed"

Discussion

The goal of this project was to develop a method for establishing the most predictive and useful Proficiency-Based benchmarks on a new VR simulator for a new Robotic Surgery platform, in this case the Medtronic HUGO™ RAS Simulator. We wanted this method to be not only robust and academically correct, but also able to predict competency and ultimately proficiency on a new robotic system by novice students. Robotic systems and robotic simulators have massive amounts of data available that can be analyzed to provide powerful insights and ultimately focused feedback to surgeons who want to either demonstrate or improve their proficiency on these complex surgical systems. In future, this kinematic data and feedback will be AI-driven.

Since we found no publications specifically describing how passing scores for each simulation exercise metric have been determined other than using "expert benchmarks," we deliberated quite a while on which of the three methodologies was the best way to define these benchmarks. There have been papers from individual centers who used their own faculty surgeons to do exercises and then determine Mean Scores for those exercise that would subsequently be applied to novices for training purposes [15, 16]. Choosing whether to set the benchmark at the expert mean, 1 SD or 2 SDs from the expert mean depends on the target training audience and the training goals. The Fundamentals of Laparoscopic Surgery (FLS) program (www.flsprogram.org) is one of the earliest examples of this [17]. In this study, the authors claimed that: "Because our goal was to design a curriculum usable by complete novices, we were concerned that setting the proficiency level at the mean expert performance levels would be overly difficult and take an unrealistic amount of time. In order to help make the proficiency levels achievable in a reasonable amount of time and to compensate for the fact that our expert group consisted of only 2 surgeons, levels for most tasks were set at 2 SDs from the determined means…." In a more recent paper, Nagaraj et al. described a similar process for setting proficiency benchmarks for another laparoscopic training curriculum: "In order to maintain high training standards, we utilized 3 unpracticed repetitions of an expert fellowship-trained minimally invasive surgeon. Expert-level performance data was then used to determine proficiency-based training benchmarks using previously reported methods: benchmark time (rounded to the nearest second) = expert mean raw score, or expert mean raw score + 1 to 2 standard deviations (SD).".[18]

Once you have established the metric benchmarks and established the scoring methodology, the final issue is to decide how many times a user must pass the exercise to be deemed proficient. We know, for example, that in the Fundamentals of Laparoscopic Surgery (FLS), the passing rate was set at 10 times in total [19, 20]. We looked at a study done by Culligan et al. in 2014 [21]. He established internal faculty expert benchmarks that were then used in a proficiency-based training curriculum by novice robotic surgeons. They were required to achieve five total passes of each exercise with two passes being consecutive to pass the entire curriculum. Predictive validity of this approach was then confirmed when the novice surgeons were able to demonstrate actual OR efficiency that was comparable to experts and superior to existing surgeons who had undergone typical pig-lab type training without being required to demonstrate proficiency on simulation. A recent systematic review of the tools used for robotic surgery technical skills assessment provides much more information on how these assessments were developed and the future utilization of AI in skills training and assessment [22].

In comparing the three different approaches (1. All data, 2. Removing the first session on each exercise, and 3. Removing outlier scores), we found the following: Method 2 (remove first sessions) assumes that the first attempts at a new exercise involve task familiarization, resulting in abnormally bad performance that affects the results. The data did not support this assumption. Method 3 (remove statistical outliers) assumes that there are external factors affecting some sessions (distractions, software glitches, etc.) and these are not representative of a "typical" attempt. Our data analysis did show that excluding obvious outlier sessions did improve the overall data conformity that leads to establishing more valid benchmarks.

Compared to the CLASSICAL method of averaging metric scores and which typically requires only a composite passing score of 80% or greater, the ABSOLUTE Method can identify individual skill components that are critical to becoming a high performing surgeon. If identified, these skills can be practiced until mastery is attained. The real test to determine the validity of these, or any new benchmarks, will take additional studies where novices and experts can be tested against these individual benchmarks and then correlate their scores with real outcomes in the O.R.

It is anticipated that these default benchmarks will initially be used "as is" by institutions, until further studies suggest updates are necessary. However, current simulation software does allow Institutions to modify these benchmarks as well as the proficiency criteria, such as number of required passes. Institutions can also create customized curricula and establish new pass/fail criteria for each metric, even selecting which metrics are considered "critical" to passing each exercise, based on their own benchmarking work (e.g., collecting data and recommendations from their own faculty).

Finally, all of the expert surgeons who participated in this study had extensive experience on the Intuitive Surgical da Vinci™ robot. The Hugo™ RAS is operated in a very similar fashion to the legacy da Vinci™ robot. This should result in a shortened learning curve for experienced da Vinci™ surgeons. However, with other new robotic platforms that have very different controls, the definition of expert will need to be defined based on experience and surgical outcomes of surgeons on those unique platforms who have gone through those learning curves.

Conclusions

This study presents an efficient and academically sound method of establishing VR simulation passing benchmarks for new robotic surgical systems using recognized robotic surgery super-users from an existing, similar robotic surgery platform. Our goal was to do this efficiently in a very short time frame, to analyze the data to determine which method (CLASSIC vs ABSOLUTE) was most beneficial for training students, and finally to ensure that each exercise was successfully passed by considering both Safety and Efficiency metrics. We feel strongly that this approach is necessary to ensure not only competency, but ultimately proficiency. We feel that this process can now be replicated in future on any new robotic surgical platform if "expert users" can be identified. We also wanted to provide insight into the considerations we believe are important when developing simulation scoring benchmarks, as they are not frequently discussed in the literature, to ensure that a user becomes proficient on a new robotic system,

This paper provides insights and visibility into the robotic VR simulation exercise benchmarking process and identifies the implications of using different methodologies to determine passing scores. Validation of the Hugo™ RAS benchmarks will occur as they are utilized by academic training centers to train surgeons on this new robotic platform.

Authors' contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Mark Brentnall, Christopher Simmonds and Anand Malpani. The first draft of the manuscript was written by John Lenihan, Mark Brentnall, and Christopher Simmonds. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding None.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest Mark Brentnall, Anand Malpani PhD: Full time employees of Surgical Science, Gottenberg. Sweden. John Lenihan M.D., Per Diem Consultant to Surgical Science, Gottenberg, Sweden. None of the other authors have any relevant financial or non-financial interests to disclose.

Ethics approval Since there were no human subjects involved in this study, no Ethics or IRB approval was required.

References

1. Seymour NE (2008) Simulation improves operating room performance. World J Surg 32(2):182–188
2. Thinggaard E, Bjerrum F, Strandbygaard J, Gögenur I, Konge L (2016) Ensuring competency of novice laparoscopic surgeons-exploring standard setting methods and their consequences. J Surg Educ 73(6):986–991. https://doi.org/10.1016/j.jsurg.2016.05.008. (**Epub 2016 Jun 17**)
3. https://www.fda.gov/medical-devices/surgery-devices/computer-assisted-surgical-systems
4. Raison N, Ahmed K, Fossati N, Buffi N, Mottrie A, Dasgupta P, Van Der Poel H (2017) Competency based training in robotic surgery: benchmark scores for virtual reality robotic simulation. BJU Int 119(5):804–811. https://doi.org/10.1111/bju.13710. (**Epub 2016 Dec 9**)
5. Gallagher AG (2012) Metric-based simulation training to proficiency in medical education: what it is and how to do it. Ulster Med J 81(3):107–113
6. Dreyfus HL, Dreyfus SE, Athanasiou T (1986) Mind over machine: the power of human intuition and expertise in the era of the computer. Free Press
7. Carraccio CL, Benson BJ, Nixon LJ, Derstine PL (2008) From the educational bench to the clinical bedside: translating the dreyfus developmental model to the learning of clinical skills. Acad Med 83(8):761–767. https://doi.org/10.1097/ACM.0b013e31817eb632
8. Wohlrab K, Jelovsek JE, Myers D (2017) Incorporating simulation into gynecologic surgical training. Am J Obstet Gynecol 217(5):522–526. https://doi.org/10.1016/j.ajog.2017.05.017. (**Epub 2017 May 13**)
9. Liss MA, Abdelshehid C, Quach S (2012) Validation, correlation, and comparison of the da vinci trainer(™) and the da vinci surgical skills simulator(™) using the mimic(™)software for urologic robotic surgical education. J Endourol 26(12):1629–1634
10. Kenney P, Wszolek M, Gould J, Libertino J, Moinzadeh A (2009) Face, content, and construct validity of dV-trainer, a novel virtual reality simulator for robotic surgery. J Urol 73(6):1288–1292

11 Salmond SE, RN (2007) Taking the mystery out of research. Box and Whisker Plots. Orthop Nursing 26(1):33

12. Simmonds C, Brentnall M, Lenihan J (2021) Evaluation of a novel universal robotic surgery virtual reality simulation proficiency index that will allow comparisons of users across any virtual reality simulation curriculum. Surg Endosc 35(10):5867–5875. https://doi.org/10.1007/s00464-021-08609-5. (**Epub 2021 Jul 6**)

13. Moglia A, Ferrari V, Morelli L, Melfi F, Ferrari M, Mosca F, Cuschieri A (2014) Distribution of innate ability for surgery amongst medical students assessed by an advanced virtual reality surgical simulator. Surg Endosc 28(6):1830–1837. https://doi.org/10.1007/s00464-013-3393-6. (**Epub 2014 Jan 18**)

14 Nahm FS (2022) Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol 75(1):25–36. https://doi.org/10.4097/kja.21209. (**Epub 2022 Jan 18**)

15. Watkinson W, Raison N, Abe T, Harrison P, Khan S, Van der Poel H, Dasgupta P, Ahmed K (2018) Establishing objective benchmarks in robotic virtual reality simulation at the level of a competent surgeon using the RobotiX Mentor simulator. Postgrad Med J 94(1111):270–277. https://doi.org/10.1136/postgradmedj-2017-135351. (**Epub 2018 Mar 6**)

16. Bilgic E, Watanabe Y, Nepomnayshy D, et. Al. (2017) Multicenter proficiency benchmarks for advanced laparoscopic suturing tasks. Am J Surg 213(2):217–221. https://doi.org/10.1016/j.am.suirg.2016.07.033

17. Ritter EM, Scott DJ (2007) Design of a proficiency-based skills training curriculum for the fundamentals of laparoscopic surgery. Surg Innov 14:107. https://doi.org/10.1177/1553350607302329

18 Nagaraj MB, AbdelFattah KR, Scott DJ, Farr DE (2022) Creating a proficiency-based remote laparoscopic skills curriculum for the COVID-19 Era. J Surg Educ 79(1):229–236. https://doi.org/10.1016/j.jsurg.2021.06.020. (**Epub 2021 Jul 3**)

19. Zendejas B, Ruparel RK, Cook DA (2016) Validity evidence for the Fundamentals of Laparoscopic Surgery (FLS) program as an assessment tool: a systematic review. Surg Endosc 30(2):512–520. https://doi.org/10.1007/s00464-015-4233-7. (**Epub 2015 Jun 20**)

20. Lerner V, DeStephano C, Ulrich A, Han ES, LeClaire E, Chen CCG (2021) A systematic review of validity evidence for the fundamentals of laparoscopic surgery examination in gynecology. J Minim Invasive Gynecol 28(7):1313–1324. https://doi.org/10.1016/j.jmig.2021.04.010. (**Epub 2021 Apr 22**)

21 Culligan P, Gurshumov E, Lewis C, Priestley J, Komar J, Salamon C (2014) Predictive validity of a training protocol using a robotic surgery simulator. Female Pelvic Med Reconstr Surg 20(1):48–51. https://doi.org/10.1097/SPV.0000000000000045.PMID:24368489.FemPelMedReconSurg

22. Boal MWE, Anastasiou D, Tesfai F, Ghamrawi W, Mazomenos E, Curtis N, Collins JW, Sridhar A, Kelly J, Stoyanov D, Francis NK (2024) Evaluation of objective tools and artificial intelligence in robotic surgery technical skills assessment: a systematic review. Br J Surg. https://doi.org/10.1093/bjs/znad331

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.