EUROPEAN UROLOGY 69 (2016) 1065–1080

available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology

Platinum Priority – Review – Education

Editorial by Alberto Breda and Angelo Territo on pp. 1081–1082 of this issue

A Systematic Review of Virtual Reality Simulators for Robot-assisted Surgery

Andrea Moglia [a,*], *Vincenzo Ferrari* [a,b], *Luca Morelli* [a,c], *Mauro Ferrari* [a], *Franco Mosca* [d], *Alfred Cuschieri* [e,f]

[a] *EndoCAS, Center for Computer Assisted Surgery, University of Pisa, Pisa, Italy;* [b] *Information Engineering Department, University of Pisa, Pisa, Italy;* [c] *Multidisciplinary Center for Robotic Surgery, University Hospital of Pisa, Pisa, Italy;* [d] *Cisanello Teaching Hospital, Pisa, Italy;* [e] *Scuola Superiore Sant'Anna, Pisa, Italy;* [f] *Institute for Medical Science and Technology, University of Dundee, Dundee, UK*

Article info

Article history:
Accepted September 14, 2015

Associate Editor:
Giacomo Novara

Keywords:
da Vinci simulator
Robotic surgery curriculum
Robotic surgery simulation

Abstract

Context: No single large published randomized controlled trial (RCT) has confirmed the efficacy of virtual simulators in the acquisition of skills to the standard required for safe clinical robotic surgery. This remains the main obstacle for the adoption of these virtual simulators in surgical residency curricula.

Objective: To evaluate the level of evidence in published studies on the efficacy of training on virtual simulators for robotic surgery.

Evidence acquisition: In April 2015 a literature search was conducted on PubMed, Web of Science, Scopus, Cochrane Library, the Clinical Trials Database (US) and the Meta Register of Controlled Trials. All publications were scrutinized for relevance to the review and for assessment of the levels of evidence provided using the classification developed by the Oxford Centre for Evidence-Based Medicine.

Evidence synthesis: The publications included in the review consisted of one RCT and 28 cohort studies on validity, and seven RCTs and two cohort studies on skills transfer from virtual simulators to robot-assisted surgery. Simulators were rated good for realism (face validity) and for usefulness as a training tool (content validity). However, the studies included used various simulation training methodologies, limiting the assessment of construct validity. The review confirms the absence of any consensus on which tasks and metrics are the most effective for the da Vinci Skills Simulator and dV-Trainer, the most widely investigated systems. Although there is consensus for the RoSS simulator, this is based on only two studies on construct validity involving four exercises. One study on initial evaluation of an augmented reality module for partial nephrectomy using the dV-Trainer reported high correlation ($r = 0.8$) between in vivo porcine nephrectomy and a virtual renorrhaphy task according to the overall Global Evaluation Assessment of Robotic Surgery (GEARS) score. In one RCT on skills transfer, the experimental group outperformed the control group, with a significant difference in overall GEARS score ($p = 0.012$) during performance of urethrovesical anastomosis on an inanimate model. Only one study included assessment of a surgical procedure on real patients: subjects trained on a virtual simulator outperformed the control group following traditional training. However, besides the small numbers, this study was not randomized.

Conclusions: There is an urgent need for a large, well-designed, preferably multicenter RCT to study the efficacy of virtual simulation for acquisition competence in and safe execution of clinical robotic-assisted surgery.

Patient summary: We reviewed the literature on virtual simulators for robot-assisted surgery. Validity studies used various simulation training methodologies. It is not clear which exercises and metrics are the most effective in distinguishing different levels of experience on the da Vinci robot. There is no reported evidence of skills transfer from simulation to clinical surgery on real patients.

[*] Corresponding author. EndoCAS, University of Pisa, Edificio 102, via Paradisa 2, 56124 Pisa, Italy. Tel./ Fax: +39 050 995689.
E-mail address: andrea.moglia@endocas.org (A. Moglia).

http://dx.doi.org/10.1016/j.eururo.2015.09.021
0302-2838/© 2015 European Association of Urology. Published by Elsevier B.V. All rights reserved.

1. Introduction

There has been a steady, almost exponential increase in the number of robot-assisted laparoscopic surgery (RAS) interventions during the last decade, reaching 570 000 in 2014 for the da Vinci surgical system (Intuitive Surgical, Sunnyvale, CA, USA) [1]. The increase persists despite the number of iatrogenic injuries caused by improper use of da Vinci robots in 2013, with the company being a defendant in more than 100 individual product liability lawsuits in 2014 and accused of inadequate surgeons training in the quest to increase sales [1,2]. In 2013 the Massachusetts Board of Registration in Medicine issued recommendations on RAS credentials, stressing proof of competency and proficiency over the number of cases performed [3]. Since many hospitals cannot afford to purchase a da Vinci robot specifically for training purposes, virtual reality (VR) simulators are considered a cost-effective solution for the acquisition of basic technical skills in RAS. On the basis of limited evidence, it has been suggested that such simulators should be integrated into proficiency-based curricula for training in basic RAS skills. However, this recommendation is contingent on high-level of evidence that skills gained during training on VR simulators can transfer to the proficiency level required for safe RAS. Currently there are five VR simulators for RAS: the Surgical Education Platform (SEP; SimSurgery, Oslo, Norway), the Robotic Surgical System (RoSS; Simulated Surgical Systems, San Jose, CA, USA), the dV-Trainer (Mimic, Seattle, WA, USA), the da Vinci Skills Simulator (dVSS; Intuitive Surgical), and the recently introduced RobotiX Mentor (3D Systems, Simbionix Products, Cleveland, OH, USA).

Although many studies have been reported, the jury is out on whether the literature provides sufficient robust evidence on the ratio of skills transfer from these VR simulators to clinical RAS surgery of the level expected for manual direct laparoscopic surgery [4]. This review seeks to address this issue by evaluating the quality and level of evidence in the literature on the efficacy of VR simulators in training for RAS, including skills transfer to the proficiency level required for clinical RAS.

2. Evidence acquisition

2.1. Search strategy

Studies were identified via searches on PubMed, Web of Science, Scopus, and the Cochrane Library up to April 2015. The Clinical Trials Database (US) and the Meta Register of Controlled Trials were also searched in April 2015. Searches were limited to the English language. The search terms used were: "da Vinci simulator" OR "robotic surgery simulator" OR "robotic surgery simulation proficiency" OR "robotic surgery curriculum". We included randomized control trials (RCTs) and nonrandomized observational studies (cohort studies) on validity and skills transfer from VR simulators for RAS to clinical RAS with a da Vinci robot. Studies on construct validity had to include assessment by expert surgeons in RAS. Other criteria for inclusion were the use of metrics to measure task execution and in some cases a subjective assessment of da Vinci robot use via global rating scales such as the Global Evaluative Assessment of Robotic Skills (GEARS; Fig. 1) [5].

2.2. Data extraction and analysis

All publications identified were scrutinized for relevance to the study before inclusion [6]. Data for all articles were extracted by one author and checked by a second author using tables according to the number of participants, interventions, comparators, outcomes, and study design (PICOS), as indicated by the Systematic Review Guidance of the Centre for Reviews and Dissemination of the University of York (UK) [7]. In the case of insufficient information from retrieved articles, the corresponding authors were contacted. The level of evidence of studies was assigned by reference to the levels of evidence identified by the Oxford Centre for Evidence-based Medicine [8]. Study quality was assessed using the Cochrane Risk of Bias Assessment tool for RCTs on a number of parameters, such as methods of randomization, allocation concealment, and blinding of assessors [9].

3. Evidence synthesis

The review is based on 36 reports (26 cohort, two case series, and eight RCTs) for 1249 participants: 28 cohort studies and one RCT on simulators validity, and two cohort studies and seven RCTs on skills transfer from VR simulators to RAS [10–45]. Two RCTs were on both validity and skills transfer [13,34]. Owing to the paucity of data on face and content validity for the RoSS simulator, two cases series (level of evidence 4) were included [26,30]. The level of evidence was 2 for RCTs and 3 for cohort studies.

3.1. Surgical specialties of participants

In validation studies, participants were recruited from the following specialties: urology ($n = 15$ studies) [10–13,15,16,21,25–27,29,30,32,36,37]; gynecology ($n = 1$) [19]; urology and general surgery ($n = 1$) [24]; gynecology and general surgery ($n = 1$) [38]; urology, gynecology, and general surgery ($n = 3$) [20,22,31]; urology, gynecology, and cardiothoracic surgery ($n = 1$) [14]; and urology, otorhinolaryngology, cardiology, thoracic surgery, and gynecology ($n = 1$) [23]; in six studies, the specialty was not indicated [17,18,28,33–35]. In skills transfer studies, participants were from gynecology ($n = 3$) [42–44]; urology ($n = 2$) [13,27]; general surgery ($n = 1$) [40]; and urology and gynecology ($n = 1$) [41]; in two studies the specialty was not reported [39,45].

3.2. Design details and quality assessment

In face and content validity studies, participants answered a questionnaire after trying the VR simulator [10–29]. For construct and discriminant validity, built-in algorithm software was used to compare scores for subjects [10–12,14–17,19–24,28–33,35,38], except for one study that used

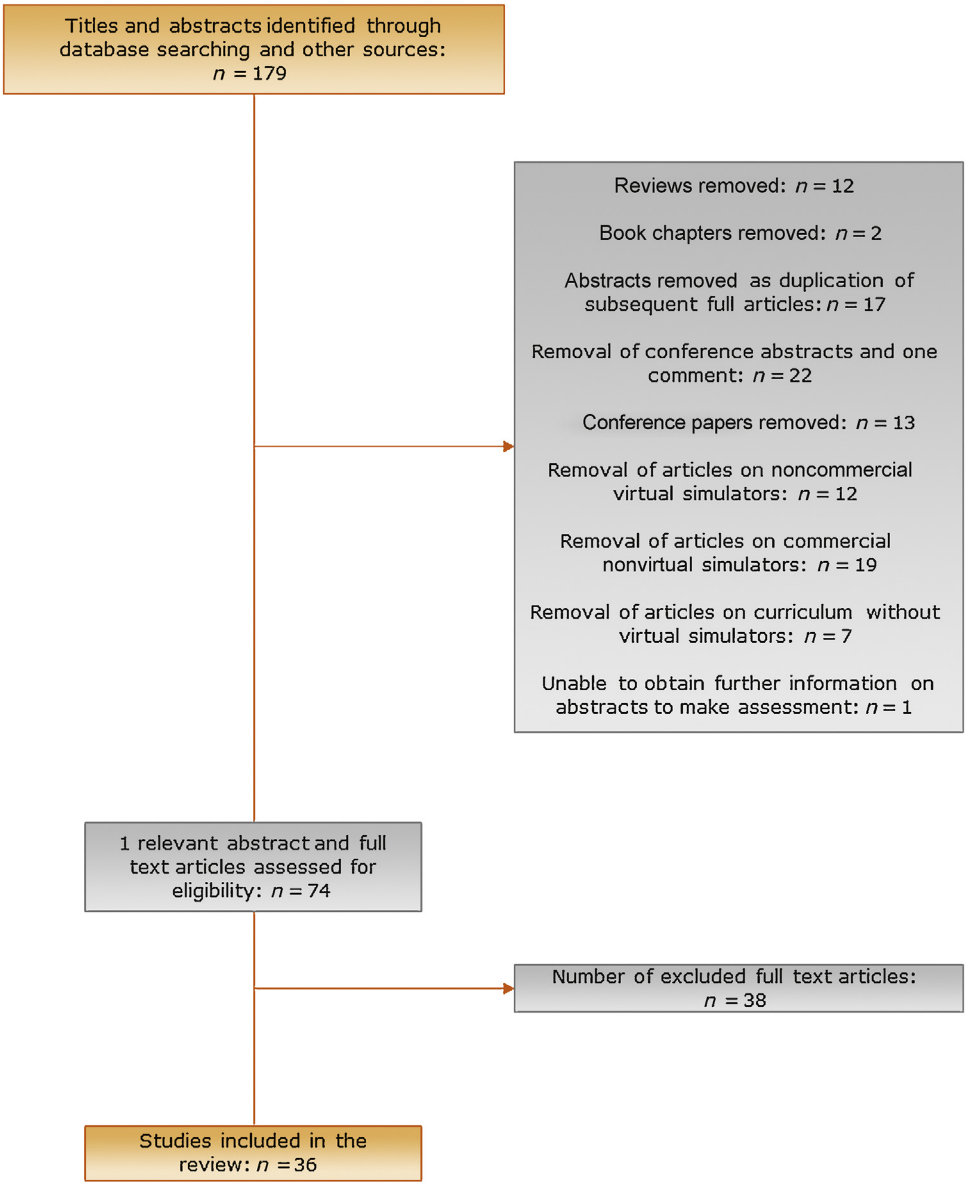
Titles and abstracts identified through database searching and other sources: n = 179
Reviews removed: n = 12
Book chapters removed: n = 2
Abstracts removed as duplication of subsequent full articles: n = 17
Removal of conference abstracts and one comment: n = 22
Conference papers removed: n = 13
Removal of articles on noncommercial virtual simulators: n = 12
Removal of articles on commercial nonvirtual simulators: n = 19
Removal of articles on curriculum without virtual simulators: n = 7
Unable to obtain further information on abstracts to make assessment: n = 1
1 relevant abstract and full text articles assessed for eligibility: n = 74
Number of excluded full text articles: n = 38
Studies included in the review: n = 36

Fig. 1 – Flow chart of the study selection process according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement [6].

an ad hoc evaluation score [34]. In concurrent validity studies, comparison between the VR simulator and robot was for the same tasks in four studies, between simulated and in vivo robotic partial nephrectomy in one study, and between simulated tasks and ex vivo animal tissue exercises in another study [15,17,18,21,25,37]. In one study on predictive validity, VR simulator performance was compared with tests using a robot on ex vivo animal tissue [37].

Five RCTs and one cohort study on skills transfer compared subjects who trained on a simulator with a control group following conventional training [27,40,42–45]. One RCT and one cohort study compared the performance of subjects who trained on a simulator with their performance on a robot [39,41]. One RCT compared subjects who trained on a simulator to robot and conventional training [13]. Two RCTs based the final assessment on execution of a procedure on an inanimate task (urethrovesical anastomosis, UA) and an animal model (cystotomy) [27,41]. Final evaluation on real patients was reported in one cohort study [43]. In other reports, final evaluation was the same as the initial evaluation on inanimate tasks [13,39,40,44,45] and animal tissue [42].

One RCT on both concurrent and predictive validity had adequate sequence generation, an unclear method for allocation concealment, and blinded assessors [37]. In the other studies on concurrent validity, assessors were blinded in three comparative studies [18,21,25] and not blinded in one study [15]; in another study the blinding was unclear [17].

For skills transfer, adequate sequence generation was reported in four RCTs, but was unclear in three [13,27,42,45]. Sealed envelopes were used for allocation concealment in three RCTs [41,42,44]. Assessors were blinded to participant identity in six RCTs, and unspecified in one [40].

3.3. *Validity studies*

3.3.1. *Face validity*

Face validity is used by experts to assess whether a test measures what it is intended to [46]. When applied to surgical simulators, it equates with realism. Twenty cohort studies reported on face validity: 12 dV-Trainer, four dVSS, two RoSS, and two SEP studies [10–29]. In the studies included in the review, apart from one report [11], face validity was evaluated by all participants including novices (without any experience on a real da Vinci console). Face validity was rated exclusively by experts in only three studies [20,21,25].

The studies used different questionnaires to assess simulator realism. Two dV-Trainer studies using the same multiple questions on a 5-item Likert scale to assess realism reported similar mean results for ease of use (3.9 vs 4.4), realism of exercises (3.9 vs 3.9), visual realism of the simulator (4.1 vs 3.6), hardware realism (3.8 vs 4.0), realistic movement (3.8 vs 3.9), and movement precision (3.1 vs 3.7) [12,18]. Similar mean results for visual realism (4.3 ± 0.8) and hardware (4.1 ± 0.7) were reported in another study that also assessed depth perception (4.0 ± 1.1), interaction with objects (4.2 ± 0.8), and instrument movements (4.1 ± 0.8) [19]. In another study that used visual analog scales, realism was rated 8 out of 10 [14]. In five reports that used an unspecified questionnaire, realism was rated high [11,13, 15–17]. In one of these comparing the dVSS and dV-Trainer, the dVSS was judged more realistic and with better hardware (foot control, 3D view, movement of masters) than the dV-Trainer ($p < 0.001$) [16]. In one study using a 6-item Likert scale, realism was rated 3.75 out of 6 by novices, and 5.11 out of 6 by experts [10]. dVSS realism was rated high on visual analog scales in two studies: 8 out of 10 in one, and 4.1 out of 5 by novices, and 4.3 out of 5 by experts in the other [23,25]. In the remaining two reports that used a Likert scale, the dVSS scored 8 out of 10 for realism, 3D graphics, and instrument control in one study, and 4.4 out of 5 in the other [22,24]. In one case series on the RoSS, 52% of participants rated the simulator somewhat close and 45% very close overall to the da Vinci robot console [26]. In one study on the SEP, software realism was rated 3.16 out of 5 by novices, and 3.4 by experts, while hardware realism was rated 3.3 by novices and 2.6 by experts (Likert scale) [28]. The other SEP study reported 3.7 out of 5 (Likert scale) for graphics realism [29].

3.3.2. *Content validity*

Content validity measures whether skills training on a simulator is appropriate and correct, that is, whether the simulator is useful as a training tool [46]. Our search identified 14 cohort studies: 10 dV-Trainer, three dVSS, one RoSS, and one SEP study [11,12,14–18,20–25,29,30]. Questions addressed by studies concerned the usefulness of simulators for training and their integration in residency programs. A problem, however, is that several studies included RAS novices in assessing the content validity of the dV-Trainer [15,17,18], dVSS [24], and SEP [29]. Content validity was assessed only by experts in five dV-Trainer studies [11,12,14,20,21] and one dVSS study [25]. Using an unspecified questionnaire, expert surgeons ranked the dV-Trainer as useful for training residents and agreed with its incorporation in the residency curriculum [11]. Surgeons also rated the dV-Trainer as a very useful training tool for residents (median 10 out of 10 on a visual analog scale) [14]. Expert robotic surgeons rated the dVSS as a very useful training system for residents (median score 8 out of 10 on a visual analog scale) [25]. Among 31 experts participating in a survey, 94% found the RoSS useful for training residents or medical students [30]. The SEP was considered useful for training by 87% of participants in one study [29].

3.3.3. *Construct validity*

Construct validity denotes the ability of a simulator to differentiate performance between experts and novices on given tasks [22]. It is important because it provides clinically meaningful assessment [46]. Our review identified 21 cohort studies on construct validity [10–12, 14–17,19–24,28–36]. Tasks were executed once in 14 studies [10,11,16,17,20–24,28,29,32,34,35], twice in one [19], three times in four [12,14,15,33], and at least twice in another study [36], as summarized in Table 1. One criticism of these studies is the small number of participants: median 32 (range 15–75) for the dV-Trainer, 38.5 (24–49) for the dVSS, 44 (27–61) for the RoSS, and 16 (12–30) for the SEP. The resulting high interstudy variability makes valid comparisons difficult. In addition, the enrolled cohorts differ among studies, with the majority comparing two groups with differing experience: novices and experts [10–12,15,20,28,29,32–36]. Other studies enrolled three groups: novices, intermediates, and experts [14,19,21–24,31]; two studies, one on the dV-Trainer and another comparing the dV-Trainer with the dVSS, included five groups [16,17].

Another issue is related to the absence of an agreed uniform method for grading the experience of participants involved in the studies. Some studies enrolled subjects without any RAS experience as novices [10,14,16,17,19–23,29,33–36]. Others used heterogeneous groups: medical students and attending surgeons, or students, residents, and specialists [16,17,19,22,23,35]. Conversely, some studies enrolled individuals with limited RAS experience in terms of the number of clinical cases performed as novices [11,12,15,24,28,31,32].

Moreover, there is no agreed definition on what constitutes an expert in RAS. Four studies rated experts according to the mean number of RAS cases performed, which ranged from 140 to 315 [11,14,17,22–24,29]. In three other studies, expert rating on RAS was based on the minimum number of cases performed, with a wide range from 30 to 240 [12,19,21,31–36]. Three other studies used the number of RAS cases performed per year or the hours spent on a da Vinci console as an index of experience [10,15,16].

The third problem is that there is no agreed definite information on which tasks confer construct validity. This

Table 1 – Level of evidence (LOE) according to the Oxford Centre for Evidence-based Medicine [8] in studies on construct validity [a]

Study	LOE	Simulator	Participants		Intervention	Assessment	Results
			n	Type and experience			
Lendvay et al [10]	3	dV-Trainer	15	11 novices (residents, 0 cases); 4 experts (surgeons: 2 <10 cases/yr, 1 10–25 cases/yr, 1 >25 cases/yr)	1 task (PB1)	6 metrics	Experts performed significantly better than novices ($p < 0.05$) for completion time, economy of motion and master control out of center
Kenney et al [11]	3	dV-Trainer	26	19 novices (students, residents, and consultant surgeons, 1.3 ± 2.2 h); 7 experts (mean 140 cases, range 30–320)	4 tasks (DN, SP, PP, PB1)	Overall score (pooled data) and 9 metrics (pooled data)	Experts performed significantly better than novices ($p < 0.05$) on overall score for all tasks and metrics except maximal force
Sethi et al [12]	3	dV-Trainer	20	15 novices (medical students and residents >1 interaction with robot); 5 experts (surgeons > 50 cases)	3 tasks (RC, SW, LB) three times	2 metrics (each task)	Experts performed significantly better than novices ($p < 0.05$) only on SW task (completion time and instruments out of view)
Hung et al [14]	3	dV-Trainer	63	16 novices (medical students, 0 cases); 32 intermediates (residents, fellows and surgeons, 0–50 cases); 15 experts (mean 315 cases, range 0–800)	10 tasks (PB2, CT2, RW2, MB2, RR2, ED1, NT, SS3, DN1, T) three times	Overall score (pooled data and each task) and 11 metrics (pooled data)	Statistically significant difference ($p < 0.05$) across all three groups for all tasks on overall score and five metrics (completion time, economy of motion, excessive instrument force, instrument collisions, and number of missed targets)
Lee et al [15]	3	dV-Trainer	20	13 novices (residents, fellows, and surgeons <50 h); 7 experts (>50 h)	4 tasks (PB2, MB1, RR1, TR) 3 times	2 metrics (each task)	Experts performed better than novices in all tasks, with significant difference ($p < 0.05$) for error in three tasks (PB2, MB1, RR1) and for completion time in TR
Liss et al [16]	3	dV-Trainer vs dVSS	32	7 medical students (0 h), 7 attending urologists (0 h), 7 junior residents (<5 h), 6 senior residents (5–50 h); 6 fellowship-trained urologists (>50 h)	1 task (T) first on dVSS then on dV-Trainer	Overall score and the same 7 metrics for dVSS and dV-Trainer (unpublished data)	Both simulators were able to differentiate experience levels among the groups (overall score, $p < 0.05$); significant difference on metrics ($p < 0.05$): dV-Trainer among the 5 groups on critical errors, economy of motion, and missed targets, for dVSS on completion time, economy of motion, and instrument collisions

Perrenot et al [17]	3	dV-Trainer	75	19 nurses and students; 37 surgeons and residents; 8 novices (0 cases);
				6 intermediates (surgeons 21 ± 12 cases);
				5 experts (surgeons 264 ± 164 cases)
					5 tasks (PP, PB1, CT1, MB1, RR1)	Overall score (pooled data) and 7 metrics (unpublished data)	Robotic surgeons (experts and intermediates) outperformed all other subjects without experience; significant difference ($p < 0.05$) for all metrics on all tasks except time of excessive force and instruments out of view

Schreuder et al [19]	3	dV-Trainer	42	15 novices (9 students, 3 residents, 3 specialists; 0 cases); 14 intermediates (2 residents and 12 surgeons; 24 cases, range 6–50); 13 experts (>240 cases, range 70–1200)	3 tasks (PB2, CT2, TR), 2 times	8 metrics (each task)	Significant difference ($p < 0.05$) during the second attempt: intermediates vs novices (PB2, completion time; CT2, economy of motion and errors; TR, completion time, economy of motion, number of instrument collisions, and errors); experts vs intermediates (PB2 and CT2, completion time and economy of motion); experts vs novices (PB2, completion time, economy of motion, and number of drops; CT2, completion time, economy of motion, master workspace range, instrument collisions, and number of drops; TR, completion time, economy of motion, number of instrument collisions, and errors)
Kang et al [20]	3	dV-Trainer	20	10 novices (residents, 0 cases); 10 experts (10–313 cases).	1 task (T3)	Overall score and 7 metrics	Experts performed significantly better than novices in overall score and all metrics ($p < 0.05$)

Hung et al [21]	3	dV-Trainer	42	15 novices (medical students, 0 cases);
				13 intermediates (surgeons <100 cases);
				14 experts (>100 cases)
					1 procedure (RPN)	10 metrics (one task)	Intermediates outperformed novices for all metrics (significant difference at $p < 0.05$ except for pierce distance error and time instruments out of view); experts scored better than intermediates for all metrics with no significant difference; experts outperformed novices for all metrics (significant difference at $p < 0.05$ except for pierce distance error and time instruments out of view)

Kelly et al [23]	3	dVSS	38	19 novices (1 resident and 18 students, 0 cases); 9 intermediates (6 residents, 1 fellow, 2 faculty member, mean 29.2 cases); 10 experts (2 residents, 1 fellow, 7 faculty members, mean 233.4 cases)	5 tasks (CT1, RR1, ES1, TR, DN1)	Overall score (each task)	Experts scored better than novices (overall score for all tasks except CT1), intermediates better than novices (overall score only for RR1, TR, and DN1, $p < 0.05$); experts scored better than intermediates (all tasks except CT1) but the difference was not significant

Table 1 (*Continued*)

Study	LOE	Simulator	Participants		Intervention	Assessment	Results
			n	Type and experience			
Finnegan et al [31]	3	dVSS	39	18 novices (residents, fellows, and surgeons, 0–20 cases); 8 intermediates (surgeons 21–150 cases); 13 experts (>150 cases)	24 tasks (all basic tasks except ED2 and SS3)	Overall score (pooled data) and 1 metric (unpublished data)	Significant difference ($p < 0.05$) on overall score in the following: intermediates vs novices, all tasks except PP, CT1, SC, ES1, ED1, SS2, and DN2; experts vs intermediates, ES1, ED1, TR, SS2, DN2, and T Significant difference ($p < 0.05$) on completion time: novices vs intermediates, PB1, PB2, RW1, RW2, RW3, MB1, RR1, RR2, ED2, NT, SS1, and DN1; intermediates vs experts, all tasks except PP, PB1, PB2, RW1, RW3, RR1, SS1, and DN1
Alzahrani et al [22]	3	dVSS	48	30 novices (1 attending, 1 fellow, 13 residents, 13 students, 2 research assistants, 0 cases); 12 intermediates (4 attending, 3 fellows, 5 residents, mean 9 cases, range 20–45); 6 experts (mean 250 cases, range 5–390)	9 tasks (PB2, RW3, MB2, RR2, ED1, NT, SS2, DN1, T)	Overall score (each task) and 11 metrics (pooled data)	Overall score: intermediates significantly better than novices ($p < 0.05$) in all tasks except RW3 and ED1; experts significantly better than intermediates in all tasks except MB2 and ED1; experts significantly better than novices in all 9 tasks Metrics: significant difference ($p < 0.05$) for intermediates vs novices for completion time, economy of motion, time of excessive force, number of instrument collisions, number of missed targets; for experts vs intermediates for master workspace range; for experts vs novices for completion time, economy of motion, excessive force time, number of instrument collisions, master workspace range, and missed targets
Lyons et al [24]	3	dVSS	46	25 novices (residents <10 cases, range 0–10); 8 intermediates (surgeons, mean 38 cases, range 12–50); 13 experts (mean 150 cases, range 60–1400)	8 tasks (PB1, PB2, RW3, MB3, RR2, ES1, SS3, T)	Overall score (each task) and up to 10 metrics (each task)	Overall score: intermediates significantly better than novices ($p < 0.05$) for all tasks except PB1 and MB3; experts significantly better than novices for all tasks; experts better than intermediates for PB1, MB3, RW3, SS3, ES1, but the differences were not significant Significant difference ($p < 0.05$) in metrics: novices vs intermediates for economy of motion for PB1, PB2, MB3, RR2, SS3, and T; completion time for PB2, MB3, RR2, ES1, SS3, and T; excessive force time for MB3, RR2, RW3, and SS3; novices vs experts for completion time and economy of motion in all 8 tasks; master workspace range in all tasks except RW3; number of instrument collisions in all tasks except MB3 and ES1; critical errors in all tasks except PB1 and ES1); intermediates vs experts for number of instrument collisions in PB1 and SS3
Hung et al [32]	3	dVSS	49	38 novices (residents <30 cases, range 0–20); 11 experts (>30 cases, range 30–2000)	4 tasks (PB2, RR2, SS3, T)	Overall score (unpublished data)	Experts had significantly better overall scores than novices ($p < 0.001$) for all tasks
Connolly et al [33]	3	dVSS	24	20 novices (medical students, 0 cases); 4 experts (>20 cases)	5 tasks (PP, PB2, CT2, MB2, SS3), 3 times	Overall score (each task) and 1 metric (each task)	Experts significantly outperformed novices ($p < 0.05$) on overall score and time in all tasks
Chowriappa et al [34]	3	RoSS	27	15 novices (surgeons, 0 cases); 12 experts (>150 cases)	4 tasks (BP, CTC, FAM, and NHE)	Overall score (each task) and 5 metrics (each task)	Experts significantly outperformed novices ($p < 0.05$) in overall score and the following metrics: all except critical errors for FAM; all except economy for CTC; all except bimanual dexterity and task time for BP; bimanual dexterity and task time for FAM; all except critical errors for NHE
Raza et el [35]	3	RoSS	61	49 novices (medical students, residents, surgeons, 0 cases); 12 experts (>150 cases)	4 tasks (BP, CTC, FAM, NHE)	Up to 10 metrics (each task)	Experts significantly outperformed novices ($p < 0.05$) for all metrics for BP; all except right tool out of view for CTC; all except left tool out of view, right tool out of view, tissue damage, and distance by camera for FAM; all except left tool out of view, right tool out of view, tool-tool collision, and number of errors for NHE

Balasundaram et al [36]	3	SEP	12	10 novices (residents, 0 cases); 2 experts (>50 cases)	5 tasks (NM, ST, SWT, ASK, IS), 10 times by residents, 2 times by experts	3 metrics (each task)	Experts significantly outperformed novices ($p < 0.05$) during second attempts in all tasks for completion time, but not path length
Van Der Meijden et al [28]	3	SEP	16	9 novices (surgeons <50 cases); 7 experts (surgeons >50 robotic surgery and laparoscopy cases)	1 task (suture)	2 metrics	Experts scored significantly better than novices ($p < 0.05$) for tool tip trajectory, but not for completion time
Gavazzi et al [29]	3	SEP	30	18 novices (students, 0 cases); 12 experts (surgeons, mean 148 cases, range 30–500)	2 tasks (AM and SK)	4 metrics (each task)	Experts significantly outperformed novices ($p < 0.05$) in both tasks on completion time, lost arrows, tool collision sum, and close entry sum

dVSS = da Vinci Skills Simulator.

[a] The 26 basic exercises on the dV-Trainer and dVSS are as follows: PP = pick and place; PB1 = peg board 1; PB2 = peg board 2; CT1 = camera targeting 1; CT2 = camera targeting 2; SC = scaling; RW1 = ring walk 1; RW2 = ring walk 2; RW3 = ring walk 3; MB1 = match board 1; MB2 = match board 2; MB3 = match board 3; RR1 = ring and rail 1; RR2 = ring and rail 2; ES1 = energy switching 1; ES2 = energy switching 2; ED1 = energy dissection 1; ED2 = energy dissection 2; NT = needle targeting; TR = thread the rings; SP1 = suture sponge 1; SP2 = suture sponge 2; SP3 = suture sponge 3; DN1 = dots and needles 1; DN2 = dots and needles 2; and T = tubes.

The other dV-Trainer basic exercises are PP_O = pick and place old; PB3 = peg board 3; T2 = tubes 2; and T3 = tubes 3

Basic dV-Trainer exercises available only for the early releases are SW = string walk; LB = letter board; RC = ring cone, Sp = suture sponge, and DN = dots and numbers.

Basic RoSS exercises: PP = pick and place; BP = ball placement; BD = ball drop; SC = spatial control; CTC = coordinated tool control; FAM = fourth arm manipulation; NHE = needle handling and exchange; CC = clutch control; NR = needle removal; FAR = fourth arm removal; TR = tissue retraction; KT = knot tying.

Basic SEP exercises: AM = arrow manipulation; NM = needle manipulation; ST = suturing without traction; SWT = suturing with traction; IS = interrupted suture; ASK = abstract square knot; SK = surgeon's knot.

holds especially true for the dV-Trainer and dVSS, which share a core set of 26 tasks developed by Mimic for basic skills training, with the number of exercises assessed varying from one to 24 (Table 1). This cannot be overlooked, since more than 90% of VR simulators installed (out of 2000) run on this common platform. Although there are fewer studies on core tasks, the number of tasks assessed is higher for the dVSS (median 5, range 1–24) than for the dV-Trainer (median 3, range 1–10) [10,11,14–17,19–24,31–33]. Studies on both simulators tend to include a small number of tasks (up to 5 tasks). There are only a few studies with a substantial number of exercises: one on the dV-Trainer (10 tasks), and three on the dVSS (8, 9, and 24 tasks) [14,22,24,31]. The number of studies evaluating tasks on the dV-Trainer and dVSS is shown in Figure 2.

The fourth issue is the variable assessment methods used. In one study on the dV-Trainer and one on the dVSS [17,31] assessment was on pooled data for overall score. In one study on the dV-Trainer, data were pooled for both overall score and metrics [11]. Pooling data on overall scores or metrics does not give an accurate estimation of the ability of a simulator to distinguish between levels of experience, since it inevitably introduces bias by mixing results for tasks of different levels of difficulty. Thus, it is not possible to recognize tasks and metrics that confer construct validity from these studies [11,17,31]. In one study on the dV-Trainer (10 tasks) and one on the dVSS (9 tasks), results for the overall score were analyzed for each task, while data for metrics were pooled [14,22]. There is only one study on the dV-Trainer (1 task), two on the dVSS (8 and 5 tasks), and one on the RoSS (4 tasks) assessment on each task on both overall score and metrics [20,24,33,34]. Three studies (one on the dV-Trainer and two on the dVSS) shared three tasks and enrolled a similar number of participants (Table 2) [14,22,24]. Only the dV-Trainer study was able to distinguish novices, intermediates, and experts, with significant difference across the three groups for all tasks ($p < 0.001$) [14]. Contrasting results were observed in the dVSS studies [22,24]. The experts achieved the highest scores only in the first task, followed by the intermediates and novices; in the other two exercises, the intermediates outperformed the experts in one study [24]. For the first task, there were significant differences between novices and experts ($p = 0.01$ and $p = 0.001$ for the first and second study, respectively) and between novices and intermediates ($p = 0.01$ and $p = 0.006$), but not between intermediates and experts ($p = 0.71$ and $p = 0.885$). For the second task, there were significant differences between novices and experts ($p < 0.01$ and $p < 0.001$) and between novices and intermediates ($p = 0.01$ and $p < 0.001$), but not between intermediates and experts ($p = 0.34$ and $p = 0.096$). For the last task, there were significant difference between novices and experts ($p < 0.01$ and $p<0.001$) and between novices and intermediates ($p = 0.01$ and $p = 0.015$), but not between intermediates and experts ($p = 0.19$ and $p = 0.447$) [22,24].

Two RoSS studies assessed construct validity for the same four tasks from the 16 tasks of the Fundamental Skills of Robotic Surgery, a curriculum based on VR simulation [34,35]. The first study used the Robotic Skills Assessment

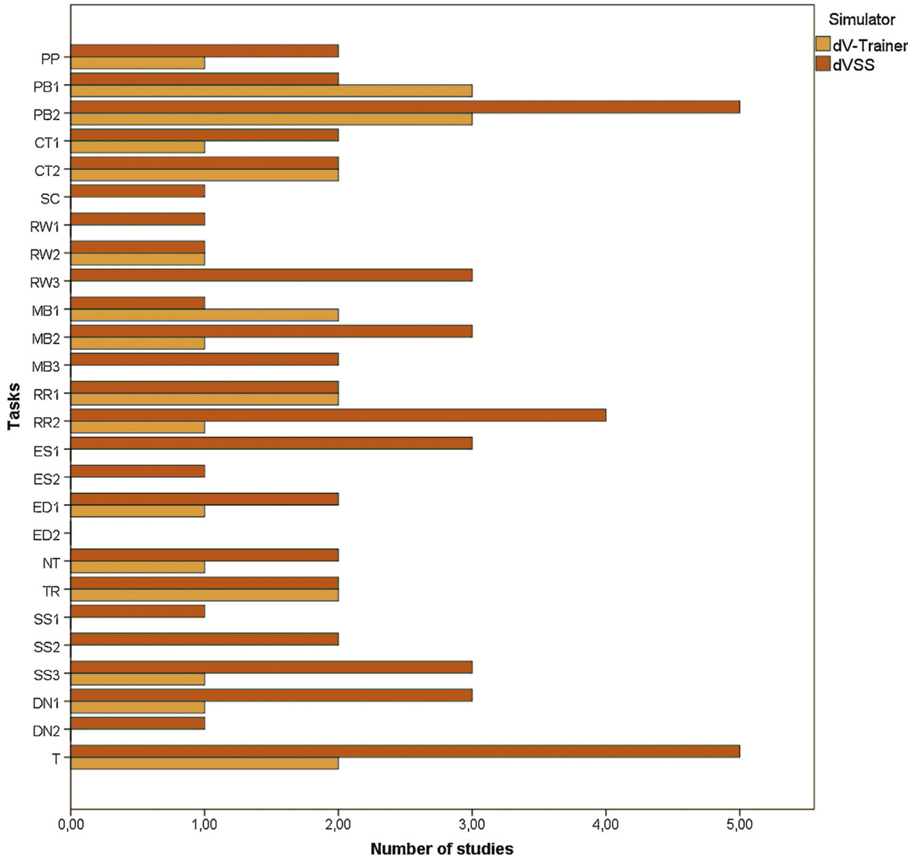

Fig. 2 – Core exercises available for the dV-Trainer and da Vinci Skills Simulator (dVSS). PP = pick and place; PB1 = peg board 1; PB2 = peg board 2; CT1 = camera targeting 1; CT2 = camera targeting 2; SC = scaling; RW1 = ring walk 1; RW2 = ring walk 2; RW3 = ring walk 3; MB1 = match board 1; MB2 = match board 2; MB3 = match board 3; RR1 = ring and rail 1; RR2 = ring and rail 2; ES1 = energy switching 1; ES2 = energy switching 2; ED1 = energy dissection 1; ED2 = energy dissection 2; NT = needle targeting; TR = thread the rings; SP1 = suture sponge 1; SP2 = suture sponge 2; SP3 = suture sponge 3; DN1 = dots and needles 1; DN2 = dots and needles 2; T = tubes.

Score (RSAS), an evaluation tool based on task time, safety in the operative field, economy, critical error, and bimanual dexterity. Experts outperformed novices on the RSAS with significant differences for all four tasks: 3.5 ± 0.1 versus 3.1 ± 0.1 ($p = 0.002$) for fourth arm control; 3.5 ± 0.1 versus 2.5 ± 0.1 ($p < 0.001$) for coordinated tool control; 3.5 ± 0.1 versus 2.1 ± 0.2 ($p < 0.001$) for ball placement; and 3.5 ± 0.1 versus 2.5 ± 0.1 ($p < 0.001$) for needle handling and exchange [34]. In the second study, experts outperformed novices in almost all metrics for the four tasks (Table 1). Three studies on the SEP found that experts performed better than novices, but the studies only included small numbers of subjects (median 16, range 12–30) assessed for limited sets of tasks (median 3, range 1–5) [28,29,36].

Finally, none of the studies identified metrics that differentiate between levels of experience. The Mimic software has 11 metrics, depending on the task, for which execution (completion) time is the most commonly used, (9 dV-Trainer and 3 dVSS studies). Although, together with economy of motion, completion time is considered as an

Table 2 – Study results reported for the same tasks

Study	Simulator	Task	Novices	Intermediates	Experts
Hung et al [14]	dV-Trainer	Peg board 2	86.9 (49.6–97.2)	90.9 (38.0–99.7)	92.1 (71.4–99.7)
Median score (range)		Ring and rail 2	45.3 (14.9–85.5)	66.1 (13.3–100.0)	81.8 (28.6–99.8)
		Tubes	56.9 (21.8–87.9)	72.2 (16.6–96.8)	84.0 (40.2–98.8)
Alzahrani et al [22]	dVSS	Peg board 2	79 (31–98)	91.5 (76–99)	91.5 (85–89)
Median score (range)		Ring and rail 2	40 (8–75)	63 (32–91)	71 (53–87)
		Tubes	50 (28–79)	67 (42–99)	78 (53–83)
Lyons et al [24]	dVSS	Peg board 2	82.3 (66.42–83.62)	94.12 (87.70–97.78)	92.76 (88.65–95.78)
Median score		Ring and rail 2	42.62 (33.8–49.95)	79.34 (65.94–86.32)	74.83 (62.62–79.93)
(95% CI)		Tubes	50.00 (44.50–56.44)	73.08 (54.44–79.75)	70.86 (63.22–78.49)

dVSS = da Vinci Skills Simulator; CI = confidence interval.

Table 3 – Study results using pooled data for the same metrics

Study	Simulator	Task	Median score (range)		
			Novices	Intermediates	Experts
Hung et al [14]	dV-Trainer	Completion time × 3	7534 (6018–10 282)	6169 (3272–15 902)	3612 (2787–6597)
		Economy of motion × 3	10 554.5 (8190.1–14 060.5)	8250.2 (6542.8–24 388.9)	6983.5 (6023.0–8845.4)
		Number of instrument collisions × 3	117 (48–262)	52 (21–349)	26 (7–54)
Alzahrani et al [22]	dVSS	Completion time × 1	2801 (1400–5823)	1795 (1258–3777)	1334 (1122–2327)
		Economy of motion × 1	3357 (2352–9851)	2580 (1987–3546)	2178 (2160–2959)
		Number of instrument collisions × 1	56 (23–180)	25 (12–80)	25 (10–37)

dVSS = da Vinci Skills Simulator.

index of surgical technical ability, it is not the ultimate measure of surgical performance and does not provide any indication of the quality of tasks executed [47]. Hence, it cannot be used as the only metric benchmark for assessment of performance on simulators, and certainly not to differentiate levels of experience. For this reason, two dVSS studies [31,33] that assessed overall score and completion time did not assess construct validity.

Two studies with pooled data for metrics on ten dV-Trainer tasks and nine dVSS tasks [14,22] are comparable since they enrolled similar numbers of participants of similar experience, shared most tasks, and evaluated the same 11 metrics (Table 3). Exercises were performed three times in the first study and once in the second [14,22]. The first study reported statistically significant differences in completion time, economy of motion, and number of instruments collisions across the three groups ($p < 0.05$) [14]. In the second study there were significant differences for all three metrics between novices and experts ($p < 0.01$) and between novices and intermediates ($p < 0.01$), but not between intermediates and experts (completion time $p = 0.17$; economy of motion $p = 0.53$; number of instrument collisions $p = 0.56$) [22], which confirms that data pooling is not appropriate for valid assessment.

3.3.4. Concurrent validity

Concurrent validity is a measure that reflects whether test scores between different instruments or simulators are in broad agreement [46]. Concurrent validity was reported in six studies (one dVSS RCT, and four dV-Trainer and one dVSS cohort study; Table 4) [15,17,18,21,25,37]. However, small numbers of participants were enrolled (median 31, range 12–75). The only RCT, limited to subjects with no or limited RAS experience (range 0–10 cases), showed high correlation between economy of motion and efficiency ($r = –0.5$), depth perception ($r = –0.6$), and bimanual dexterity ($r = –0.7$; $p < 0.01$), and between time of excessive force and tissue handling ($r = –0.7$, $p = 0.0002$) [37]. However, the three da Vinci robot tasks on ex vivo animal tissue required cutting and excision, not present in the 17 dVSS tasks. Hence, the two tests were improperly compared for concurrent validity [37]. Three dV-Trainer studies evaluated correlation between identical tasks on a simulator and a robot [15,17,18], but no task was shared among the three studies. Only limited concurrent validity has been reported for the dV-Trainer. One study showed no correlation between the simulator and robot for each task, but pooled data for four tasks revealed correlation for time ($r = 0.55$, $p = 0.026$) and errors ($r = 0.62$, $p = 0.011$) [15]. The second study found correlation for only two out of five tasks, and on pooled data for completion time ($r = 0.64$) and economy of motion ($r = –0.71$) [17]. The third study reported correlation for completion time for all four tasks ($r = 0.60–0.62$, depending on the task) [18].

3.3.5. Predictive validity

Predictive validity is ability to select those who will not be able to perform surgical operations well, despite training, and the reverse, that is, those who will excel [46]. The present review identified only one RCT on the dVSS (Table 4), but the study included only a small number of participants ($n = 24$) [37]. There was high correlation ($r = 0.7$, $p < 0.0001$) between the overall score during the initial test on the simulator and the final test on a da Vinci robot according to Global Operative Assessment of Laparoscopic Skills (GOALS) [37]. The study did not provide information on the individual performance of participants.

3.3.6. Discriminant validity

Discriminant validity is the capability of a simulator to differentiate ability levels within a group with similar experience [46]. The review identified two dVSS cohort studies that assessed innate surgical ability among medical students [38,48]. One of these is also the only validity study in which participants executed all 26 basic core tasks of the Mimic software (Table 4) [38]. The study evaluated psychomotor skills among 121 medical students and found that 6.6% had outstanding and 11.6% had low-level psychomotor skills [38]. There was no significant difference in overall score between top students and four expert surgeons on a da Vinci console (median 62.1, range 55.2–67.5 vs median 52.7, range 50.4–56.2; $p = 0.368$). However, the study had no follow-up to confirm the initial findings [38].

3.4. Skills transfer studies

It took 6 yr for the first publication demonstrating level 2 of evidence (in RCTs) of positive transfer of skills from the MIST VR (Mentice, Gothenburg, Sweden), the first VR simulator for laparoscopy, to surgery on patients, and

Table 4 – Level of evidence (LOE) according to the Oxford Centre for Evidence-based Medicine [8] in studies on concurrent, predictive, and discriminant validity

Study	LOE	Simulator	Participants		Intervention	Assessment	Results
			n	Type and experience			
Lee et al [15]	3	dV-Trainer	20	13 novices (residents, fellows, surgeons <50 h); 7 experts (surgeons >50 h)	4 tasks (PB2, MB1, RR1, TR), 3 times; same tasks on da Vinci robot	2 metrics	Concurrent validity: no correlation on individual 4 tasks between dV-Trainer and da Vinci robot; correlation only for pooled data for 4 tasks for time (r=0.55, p = 0.026) and errors (r=0.62, p = 0.011).
Perrenot et al [17]	3	dV-Trainer	75	19 nurses and students; 37 surgeons and residents 8 novices (0 cases); 6 intermediates (surgeons 21 ± 12 cases); 5 experts (surgeons: 264 ± 164 cases)	5 tasks (PP, PB1, CT1, MB1, RR1); same tasks on da Vinci robot	7 metrics	Concurrent validity: correlation between dV-Trainer and da Vinci robot in 2 tasks (PP r = 0.66; RR1 r = 0.62) for 2 metrics using pooled data for all tasks (time r = 0.64; economy of motion r = –0.71)
Egi et al [18]	3	dV-Trainer	12	9 intermediates and 3 experts (>50 laparoscopic procedures)	4 tasks (PP, PB1, TR, SS1) on dV-Trainer; same tasks on da Vinci robot	1 metric OSATS	Concurrent validity: correlation between tasks on dV-Trainer and da Vinci robot on time (r = 0.60–0.62 depending on task, p = 0.030–0.041) and between overall OSATS score and SS1 task (r = 0.58, p = 0.046).
Hung et al [21]	3	dV-Trainer	42	15 novices (medical students, 0 cases); 13 intermediates (surgeons <100 cases); 14 experts (surgeons >100 cases)	1 task (renorrhaphy) at simulators; 1 procedure: RPN	1 procedure (RPN); GEARS for videos of simulated task and in vivo porcine RPN with da Vinci robot	Concurrent validity (only experts and intermediates): high correlation (r = 0.8, p < 0.0001) on overall score (GEARS) between simulated renorrhaphy and in vivo RPN
Hung et al [37]	2	dVSS	24	2 medical students, 14 residents, 5 fellows, 1 intern, and 2 staff members with 0 cases (range 0–10)	17 tasks (all basic tasks except PP, PB1, CT1, RW1, MB1, RR1, ES2, SS1, SS2). Participants performed 3 ex vivo tasks with da Vinci robot on animal tissue (bowel resection, cystotomy and repair, and partial nephrectomy) and then randomized in 2 groups: group 1 trained for 10 wk with 17 tasks on dVSS; group 2 used the simulator after initial test on animal and at end of study; final evaluation of the same 3 tasks on animal tissue with a da Vinci robot	Overall score and up to 11metrics (depending on task) GOALS	Concurrent validity: high correlation on overall score between baseline on simulator and initial animal test (r = 0.7, p < 0.0001) High correlation between economy of motion and efficiency (r = –0.5), depth perception (r = –0.6), and bimanual dexterity (r = –0.7, p < 0.01), and between time of excessive force and tissue handling (r = –0.7, p = 0.0002) Moderate correlation between simulator and robot for number of instrument collisions (r = 0.5, p = 0.01) Completion time on simulator moderately correlated with time to complete animal tissue tasks (r = 0.4, p = 0.06) and autonomy (r = –0.6, p = 0.004). Predictive validity: high correlation between baseline overall score on simulator and final test on animal tissue assessed using GOALS (r = 0.7, p < 0.0001)
Ramos et al [25]	3	dVSS	36	24 novices (0 cases); 12 experts (200 cases, range 30–2015)	3 tasks on dVSS (PP, PB2, MB1); same tasks on da Vinci robot	Overall score and 7 metrics on dV-Trainer; 3 metrics on da Vinci robot; GEARS	Concurrent validity: moderate correlation (r = 0.54, p < 0.001) between overall score for dVSS and total GEARS score for da Vinci robot High correlation between time to complete on simulator and GEARS time (r = 0.87) and efficiency (r = –0.79, p < 0.001) Economy of motion: moderate to high correlation with efficiency (r = 0.64), depth perception (r = 0.75), and bimanual dexterity (r = 0.63, p < 0.001); moderate correlation for number of instrument collisions (r = –0.48, p = 0.002) and master workspace range (r = 0.48, p = 0.001)
Moglia et al [38]	3	dVSS	125	121 medical students (0 cases); 4 experts (250 cases, range 100–350)	26 tasks (all basic tasks)	Overall score and 3 metrics	Discriminant validity: significant difference (p < 0.05) on overall score, completion time, economy of motion, and time of excessive force for 6 selected tasks (CT1, RW2, RW3, RR2, NT, TR) among three different subpopulations of medical students (high talent, 6.6%; average, 81.8%; and low talent, 11.6%)

dVSS = da Vinci Skills Simulator; RPN = radical partial nephrectomy. PP = pick and place; PB1 = peg board 1; PB2 = peg board 2; PB3 = peg board 3; CT1 = camera targeting 1; CT2 = camera Targeting 2, SC = Scaling, RW1 = Ring Walk 1, RW2 = Ring Walk 2, RW3 = Ring Walk 3, MB1 = Match Board 1, MB2 = Match Board 2, MB3 = Match Board 3, RR1 = ring and Rail 1, RR2= Ring and Rail 2, ES1= Energy Switching 1, ES2= Energy Switching 2, ED1= Energy Dissection 1, ED2= Energy Dissection 2, NT= Needle Targeting, TR= Thread the Rings, SP1= Suture Sponge 1, SP2= Suture Sponge 2, SP3= Suture Sponge 3, DN1= Dots and Needles 1, DN2 = dots and needles 2; T = tubes; T2 = tubes 2; T3 = tubes 3; GOALS = Global Operative Assessment of Laparoscopic Skills; GEARS = Global Evaluation Assessment Robotic Surgery; OSATS = Objective Structured Assessment of Technical Skills.

a similar interval for the LapSim (Surgical Science, Gothenburg, Sweden), the most validated VR simulator for laparoscopy [49–52]. Initial studies on current VR simulators for robotic surgery date back to 2008 [36,53,54], but there is still no evidence of comparable level. This accounts for the general reluctance by residency program directors to accept and incorporate VR simulators for RAS as key components in surgical curricula.

To date, skills transfer has been reported on inanimate models in one cohort study and five RCTs, on animal tissue models in two RCTs, and on real patients in one cohort study [13,27,39–45]. These studies (four dV-Trainer, three dVSS, two RoSS) are summarized in Table 5 according to the Systematic Review Guidance of the Centre for Reviews and Dissemination of University of York (UK) [7]. The major weakness of these studies relates to the small numbers of participants (range 12–53). In five of seven studies on the dV-Trainer and dVSS, the experimental group trained on the VR simulator outperformed the control group (Table 5), while in the other two (both RCTs), in which the control group was trained on a real da Vinci robot, equivalent performance was reported for both groups [13,41]. The participants had to reach proficiency equivalent to that of expert surgeons once in one dV-Trainer RCT and one dVSS cohort study [41,43]. This contrasts with the criterion used for manual laparoscopic surgery, which mandates documented proficiency on two consecutive occasions [4,43]. In three RCTs (one dV-Trainer and two dVSS) proficiency was not based on expert benchmarking but on an arbitrary value: an overall score of 80% in two studies, and 60% for all metrics in the third [13,42,44]. In one RCT, assessment of simulator training was time- rather than proficiency-based [39,40]. Other studies have used a simplistic final assessment involving the same inanimate tasks as for the initial evaluation [13,39,40,45]. Only two RCTs based the final assessment on execution of a procedure: UA on an inanimate model and cystotomy on an animal (swine) model [27,41]. In the first study (RoSS), the experimental group outperformed the control group, with a significant difference for GEARS (overall score $p = 0.012$, bimanual dexterity $p = 0.016$, and force sensitivity $p < 0.001$) and for objective UA assessment (all three metrics $p < 0.05$; Table 5). The other dV-Trainer study found no significant difference between the experimental and control groups ($p < 0.05$) for time to perform cystotomy closure and overall GEARS score (Table 5).

Final evaluation on real patients was reported in only one study that included 14 subjects in the experimental group and four in the control group for hysterectomy procedures [43]. The simulator-trained group outperformed control one during hysterectomy with a da Vinci robot on overall GOALS score (34.7 vs 31.1; $p = 0.07$), with significant differences for time (21.7 ± 3.3 vs 30.9 ± 0.6 min; $p < 0.0001$) and estimated blood loss (25.4 vs 31.25 ml; $p < 0.0001$, Table 5). Besides the small numbers, this study lacked randomization.

Currently there is one ongoing RCT comparing different training modalities for robotic surgery including VR simulators. The study design includes baseline evaluation on a task specific to cardiac surgery, followed by randomization into four groups: wet laboratory (pig) training, dry laboratory training (Fundamentals of Laparoscopic Surgery tasks), dVSS training, and a control group, with the three experimental arms trained until proficiency. Final evaluation is on tasks relevant to robotic cardiac surgery. Primary outcomes include time to complete mitral valve annuloplasty and time to complete 10-cm dissection of the internal thoracic artery. Assessment is conducted using GEARS for each task [55].

3.5. *Studies on simulated procedures for urology*

Approximately 91 000 RAS urology procedures were performed in the USA in 2014 (20% of total procedures) [1]; prostatectomy was the most common urology procedure, well exceeding partial nephrectomy. In 2014, approximately 60 000 prostatectomies were executed in the USA and 65 000 in other countries [1].

One series of tasks (Tubes group) for UA, a complicated step in robot-assisted radical prostatectomy (RARP), is available for both the dV-Trainer and dVSS. Of the six reports on the basic version, four provided results (Table 1) [14,22,24,31]. The latest version (Tubes 3) is exclusive to the dV-Trainer [20]. This task was rated very realistic by experts (4.5 out of 5 on a Likert scale) and useful for training (4.3 out of 5) according to face and content validity. Experts performed better than novices, with a statistically significant difference for all metrics ($p < 0.05$) and overall score (median 240.0, range 26.0–359.5 vs median 13.8, range 11–20; $p = 0.016$) [20].

Beside basic tasks for familiarization with the da Vinci interface and controls, current VR simulators for robotic surgery offer simulated procedures for urology: VR prostatectomy (RobotiX Mentor) and augmented reality for partial nephrectomy (dV-Trainer) and prostatectomy (RoSS). In a dV-Trainer study on an augmented reality module for partial nephrectomy, the procedure was rated realistic and effective as a training tool for residents, with median of 8.0 (range 5–10) for face validity and 8.2 (range 1–10) for content validity for all modules (colon mobilization, Kocherization of the duodenum, hilar dissection, kidney mobilization, and tumor resection and repair). Experts scored better than novices on time ($p = 0.009$) and accuracy ($p = 0.004$) for anatomy exercises. Regarding technical questions, experts and novices were comparable for time ($p = 0.1$) but not accuracy ($p = 0.004$). The opposite was found for operation steps, with similar accuracy ($p = 0.3$) but not time ($p = 0.02$). There were significant differences between novices and intermediates only for technical questions ($p = 0.02$ for time and $p = 0.007$ for accuracy); between experts and intermediates the only significant difference was for questions on operation steps ($p = 0.02$ for time). The partial nephrectomy module also includes one VR task simulating renorrhaphy. Only experts and intermediates tested it, with experts performing better than intermediates for all GEARS domains and overall score (median 28, range 18–30 vs 18, range 15–23; $p = 0.002$). Concurrent validity between in vivo porcine nephrectomy and VR renorrhaphy task revealed high correlation ($r = 0.8$) for all GEARS

Table 5 – Level of evidence (LOE) according to the Oxford Centre for Evidence-based Medicine [8] in studies on skills transfer from a virtual reality simulator to a da Vinci robot

Study	LOE	Simulator	Participants (*n*)	Groups and experience	Comparator	Intervention	Results
Lerner et al [39]	3	dV-Trainer	23	dV-Trainer group (8 students, 3 urology interns, 1 urology fellow) with no robotic surgery experience; da Vinci robot group (10 urology residents and 1 urology fellow) with minimal robotic surgery experience	dV-Trainer, 4 tasks (PP, LB, RW, CC) 4 times; da Vinci robot, 5 tasks (PB, PC, LB, SR, and IKT) 4–6 times	Pretest (baseline): 5 tasks on inanimate models with da Vinci robot; Posttest: same 5 tasks on robot. Assessment: time	Similar improvements for both groups; da Vinci robot group faster on PC (282 [120–430] vs 385 (187–1409]), LB (284 [163–508] vs 316 [144–666]), and IKT (168 [103–327] vs 312 [190–687]); similar scores on PB (92 [37–170] vs 91 [48–143]) and SR (74 [36–167] vs (72 [32–158])
Korets et al [13]	2	dV-Trainer	16	dV-Trainer group (3 residents < 50 cases, and 2 > 50); da Vinci robot group (3 residents <50 cases, and 2 >50); control group (4 residents <50 cases, and 2 >50)	dV-Trainer, 15 tasks with 80% as passing score; da Vinci robot, 90 min of one-to-one training with a fellow; control, standard training	Pretest (baseline): 2 tasks on inanimate models with da Vinci robot (ring and wire, and suture); Posttest: same 2 tasks on robot Assessment: time and OSATS	Posttest: dV-Trainer and da Vinci robot scored better than control group: 202.1 ± 43.6 vs 122.8 ± 42.3 vs 236.7 ± 80.0 for time on ring wire; 138.7 ± 43.1 vs 93.2 ± 38.1 vs 147.6 ± 35.5 for time on KT; 13.0 ± 1.2 vs 14.4 ± 1.0 Vs. 11.7 ± 1.1 for overall OSATS dV-Trainer: improved time on both tasks (ring and wire $p = 0.04$; KT $p = 0.10$) and on OSATS ($p = 0.03$); da Vinci group: improved time on both tasks (ring and wire $p < 0.01$; KT $p = 0.02$) and on OSATS ($p < 0.01$) Control group: improved time on both tasks (ring and wire $p = 0.16$; KT $p = 0.14$) and on OSATS, but the latter was not significant ($p = 0.09$)
Cho et al [40]	2	dV-Trainer	12	Experimental group (6 surgeons, 40 laparoscopy cases, range 25–60); control group (6 surgeons, 55.8 cases, range 30–200)	Experimental, 3 wk with simulator curriculum; control, no use of simulator or robot	Pretest: 3 tasks on dV-Trainer (PP, PB1, TR); 2 tasks with da Vinci robot on inanimate models (needle control, and suture and tying) Posttest: same as pretest Assessment: VR index based on completion time and economy of motion DV index (completion time and accuracy)	Experimental scored significantly better than control group for VR index (19.3 ± 4.5 vs 9.7 ± 4.1, $p = 0.001$) and DV index (5.80 ± 1.13 vs 4.05 ± 1.03, $p = 0.028$)
Whitehurst et al [41]	2	dV-Trainer	20	dV-Trainer group (4 residents, 3 fellows, 3 attending); da Vinci robot (2 residents, 6 fellows, 2 attending)	dV-Trainer, 3 tasks (PP, PB1, RW1) until proficiency set by 3 expert surgeons; da Vinci, 3 FLS tasks (PT, CC, IS) until proficiency set by 3 expert surgeons	Cystotomy closure with robot on pig model Assessment of videos by GEARS	No significant difference in time to perform cystotomy closure and overall GEARS score (2.96 ± 0.77 for da Vinci group vs 2.83 ± 0.66 for dV-Trainer group)
Vaccaro et al [42]	2	dVSS	18	dVSS group (9 residents); control group (9 residents)	dVSS, standard robotic orientation and 9 tasks (RW1, RW2, RW3, ES1, ES2, ED2, SS2, SS3) with passing score of 80% and 60% for all metrics; control, standard robotic orientation	Pretest: incision and suture task with robot on animal tissue Posttest: same as pretest Assessment: total time, time to incision, and suture; GRS and rOSATS	Posttest: simulator outperformed control group on time: 14.0 (9.8–16.6) vs 24.6 (16.8–26.0) for total time ($p = 0.058$); 1.8 (1.4–2.4) vs 4.3 (3.9–6.4) for time to incision ($p = 0.042$); 11.4 (8.3–13.8) vs 18.3 (12.8–20.7) for suturing time ($p = 0.145$); no significant difference for rOSATS (15.0 ± 1.4 vs 13.3 ± 4.2, $p = 0.242$) or GRS (18.6 ± 3.1 vs 15.7 ± 5.0, $p = 0.202$)

Culligan et al [43]	3	dVSS	18	dVSS group (14 surgeons: 0 cases); control group (4 surgeons with enough cases to operate unsupervised)	dVSS, online orientation of robot, 10 tasks (PB2, CT2, RW3, MB2, MB3, ES1, ED1, ED2, SS2, T) until proficiency set by 5 expert surgeons, and pig laboratory; control, standard activities	Supracervical hysterectomy with robot Assessment: GOALS.	dVSS outperformed control group on time (21.7 ± 3.3 vs. 30.9 ± 0.6, $p < 0.0001$), estimated blood loss (25.4 vs 31.25, $p < 0.0001$), and overall GOALS score (34.7 vs 31.1), but the latter was not significant ($p = 0.07$)
Kiely et al [44]	2	dVSS	23	dVSS group (8 residents and 5 attending); control group (9 residents and 1 attending); both groups without experience at da Vinci master console as primary operator	dVSS, 5 tasks (CT1, CT2, SS1, SS2, SS3) until all metrics had green checkmark twice consecutively and suture task 10 times; control, conventional training	Pretest: 1 task on dVSS (SS1) and suture on inanimate vaginal cuff model with real da Vinci robot Posttest: same as pretest Assessment: GOALS+ and GEARS	Posttest: dVSS significantly ($p < 0.05$) outperformed control group on both tasks (overall score, GOALS+, and GEARS)
Stegemann et al [45]	2	RoSS	53	RoSS group, 30 (medical students, residents, fellows, and attending surgeons); control group, 23 (medical students, residents, fellows, and attending surgeons)	RoSS, introduction to use of da Vinci robot, 16 FSRS tasks (no proficiency required); control, introduction to use of da Vinci robot	3 tasks (BP, suture pass, and 4th arm manipulation) on inanimate models with a robot, three times Assessment of videos using ad hoc parameters	RoSS performed better than control group on 3 da Vinci robot tasks; the difference was only significant for number of slips for BP (1.5 ± 0.2 vs 2.5 ± 0.3, $p = 0.14$) and instruments out of view for suture pass task (0.5 ± 0.1 vs. 1.1 ± 0.2, $p = 0.026$)
Chowriappa et al [27]	2	RoSS	52	Experimental group, 12 residents and 14 fellows (<25 h of robotic surgery); control group, 10 residents and 16 fellows (<25 h of robotic surgery)	Experimental, 4 tasks (BP, SC, TR, KT) on RoSS, 4 sessions of augmented reality for urethrovesical anostomosis (HoST), introduction to da Vinci robot, and 4 tasks on robot (BP, suture pass, 4th arm manipulation, and suturing); control, same as experimental, but watched 4 videos instead of HoST training	Urethrovesical anastomosis on inanimate model Assessment: GEARS and objective urethrovesical anastomosis score	Experimental outperformed control group according to GEARS (overall score 14.4 ± 1.2 vs 11.9 ± 4.1, $p = 0.012$; bimanual dexterity 2.9 ± 0.2 vs 2.4 ± 1.0, $p = 0.016$; force sensitivity 2.5 ± 0.2 vs 2.0 ± 0.8, $p < 0.001$) Experimental outperformed control group for all objective urethrovesical anastomosis metrics: 3.0 ± 0.7 vs 2.4 ± 0.8 for needle position ($p = 0.008$); 3.0 ± 0.9 vs 2.3 ± 1.0 for needle driving ($p = 0.042$); 3.4 ± 0.9 vs 2.6 ± 0.9 for suture placement and tissue manipulation ($p = 0.014$)

dVSS = da Vinci Skills Simulator; LB = letter board; SR = string running; P*p* = pick and place; RW = ring walk; CC = clutching cavity; PB1 = peg board 1; PB2 = peg board 2; PC = pattern cutting; CT1 = camera targeting 1; CT2 = camera targeting 2; RW1 = ring walk 1; RW2 = ring walk 2; RW3 = ring walk 3; MB2 = match board 2; MB3 = match board 3; ES1 = energy switching 1; ES2 = energy switching 2; ED2 = energy dissection 2; SP1 = suture sponge 1; SP2 = suture sponge 2; SP3 = suture sponge 3; T = tubes; BP = ball placement; SC = spatial control; TR = tissue retraction; KT = knot tying; IKT = intracorporeal KT; PT = peg transfer; CC = circle cut; LL = ligating loop; ES = extracorporeal suture; IS = intracorporeal suture; FSRS = Fundamental Skills for Robotic Surgery; FLS = Fundamentals of Laparoscopic Surgery; GOALS = Global Operative Assessment of Laparoscopic Skills; GEARS = Global Evaluation Assessment Robotic Surgery; OSATS = Objective Structured Assessment of Technical Skills; GRS = Global Rating Scale.

domains, with overall median scores 20 (range 13–29.5) and 21 (range 15–30), respectively [21].

In one RoSS RCT, 70% of subjects found the augmented reality module realistic for UA. An experimental group trained on an augmented reality module outperformed a control group performing UA on an inanimate model, with significant differences using both GEARS (overall score 15.3 ± 3.2 vs 11.9 ± 4.1; $p = 0.008$) and objective UA evaluation (3.4 ± 0.9 vs 2.6 ± 0.9; $p = 0.014$; Table 5) [27].

3.6. *Future research*

VR surgical simulators, including those for robotic surgery, have followed the long established benefit in terms of safety obtained by training on flight simulators in well-established use by the aviation industry for several decades. The aviation industry has for years adopted competency-based curricula requiring pilots to meet specific benchmark performance criteria before moving to the next level of training. Currently, this cannot be said of RAS. Indeed, there is no standard proficiency-based curriculum for this emerging subspecialty of surgery. The validation trial on Fundamentals of Robotic Surgery (FRS), a proficiency-based curriculum involving 14 training centers accredited by the American College of Surgeons, is a step in the right direction. The Robotic Training Network is another curriculum development initiative. It involves 50 centers in the USA and aims to standardize the robotic surgical curriculum and education for residents in gynecology and general surgery. Other curricula have been proposed by the University of Pennsylvania and the University of Toronto [56–58].

The European Association of Urology Robotic Urology Section (ERUS) developed a structured curriculum that includes theoretical training, simulation training (dry lab, wet lab, and VR simulation), case observation, and a fellowship program consisting of assisting with and then performing segments of a procedure before undertaking a whole procedure (modular training, dual console) [59]. In a pilot study of 10 fellows with median experience of 4 mo on a da Vinci console, eight (80%) were considered by their mentors able to perform a RARP independently, safely, and efficiently on completion of the ERUS curriculum, and three (30%) were considered able to perform a complex RARP independently, safely, and effectively. The generic dedicated scoring criterion for each procedural RARP step showed construct validity, since two experts outperformed the fellows (mean overall score 13.6 vs 11.0). Technical skills were evaluated using four tasks on the dVSS; the overall score on the tasks improved, with a statistically significant difference from baseline ($p < 0.05$) over the training period [60].

A proficiency-based surgical curriculum for RAS has to be flexible, as it needs to cater for surgeons of different grades and surgical experience. Ideally, it should incorporate initial and final assessments before certification of proficiency in RAS. Initial baseline evaluation is essential, as innate ability varies among individuals, so some require more training than the average, whilst a few require less training to reach proficiency. Some of the current exercises can be used to test innate ability for surgery. They are by no means perfect, and manufacturers of RAS VR simulators should be encouraged to improve all systems, ideally by working closely with surgeons and experts in training and behavior science and human factor engineers [38]. A comprehensive curriculum should also include follow-up tests to assess skills retention or deterioration, sometimes referred to as revalidation. Currently little is known about skills maintenance gained on VR simulators for RAS, apart from two studies, both with small cohorts of participants [61,62]. These studies reported contrasting findings: skills deterioration for 12 tasks after reaching proficiency, and maintenance of skills, as distinct from proficiency, for six tasks [61,62].

The cost-effectiveness of VR simulators for RAS compared to other training approaches is largely unknown except for one RoSS study [63]. In a study of 105 trainees, RoSS cost-effectiveness was evaluated by computing time spent on training on the RoSS instead of using a real da Vinci robot for training. Time spent on the RoSS was 361 h, equivalent to 73 potential cases in the operating room according to the average duration for RAS procedures at the Roswell Park Center Institute. Use of the da Vinci surgical system for training instead of scheduling it for operating on real patients would have resulted in loss of 73 cases, corresponding to a loss of over $600 000 in net patient revenue, approximately five times the RoSS price ($125 000) [63]. VR simulators from other vendors are similarly priced: $158 000 for the dV-Trainer, $100 000 for the RobotiX Mentor, and $90 000$ for the dVSS (requiring a dedicated da Vinci console at an additional $500 000).

In practical terms, the transfer effectiveness ratio is the only valid measure of cost-effectiveness. The transfer effectiveness ratio is used by the aviation industry to indicate the difference in time required to achieve fully competent performance between flying a real aircraft and virtual flying on a flight simulator under various scenarios, such as poor weather conditions and engine malfunction [64]. For flight simulators, the ratio ranges from 0.67 to 0.99; that is, 1 h on a flight simulator saves approximately 40–60 min of real flying time [65]. At present there are no data on the transfer effectiveness ratio from VR simulators to clinical RAS. The question we need to answer as trainers and educators of residents is, why not.

4. Conclusions

The aim of this review was to evaluate the level of evidence in published studies on the validity and skills transfer of virtual simulators for robot-assisted surgery. The variability in study design makes comparisons difficult. Overall there is no evidence on the transfer of skills gained using virtual simulators to the operating room. For this reason large, RCTs, preferably multicenter, are required to solve this issue, with the ultimate goal of facilitating the adoption of VR simulators in curricula for robotic surgery.

Author contributions: Andrea Moglia had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Moglia, Cuschieri.
Acquisition of data: Moglia.
Analysis and interpretation of data: Moglia, Cuschieri.
Drafting of the manuscript: Moglia, Cuschieri.
Critical revision of the manuscript for important intellectual content: Moglia, Ferrari V, Morelli, Ferrari M, Mosca, Cuschieri.
Statistical analysis: Moglia.
Obtaining funding: Mosca.
Administrative, technical, or material support: Moglia.
Supervision: Cuschieri.
Other: None.

Financial disclosures: Andrea Moglia certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

Funding/Support and role of the sponsor: This study was supported by Fondazione Arpa (www.fondazionearpa.it). The sponsor played a role in the design and conduct of the study.

References

[1] Intuitive Surgical. Investor relations. http://investor.intuitivesurgical.com/phoenix.zhtml?c=122359&p=irol-irhome
[2] Bernstein Leibhard LLP. da Vinci robot lawsuit information center. www.davincirobotlawsuitcase.com
[3] Commonwealth of Massachusetts Board of Registration in Medicine. Advisory on robot-assisted surgery. www.mass.gov/eohhs/docs/borim/physicians/pca-notifications/robot-assisted-surgery.pdf
[4] Dawe SR, Windsor JA, Broeders JA, et al. A systematic review of surgical skills transfer after simulation-based training: laparoscopic cholecystectomy and endoscopy. Ann Surg 2014;259:236–48.
[5] Goh AC, Goldfarb DW, Sander JC, et al. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. Urology 2012;187:247–52.
[6] Transparent Reporting of Systematic Reviews and Meta-analyses. The PRISMA statement. www.prisma-statement.org/statement.htm
[7] Centre for Reviews and Dissemination. Guidance for undertaking reviews in health care. www.york.ac.uk/crd/guidance
[8] Oxford Centre for Evidence-based Medicine. Levels of evidence 2011. www.cebm.net/wp-content/uploads/2014/06/CEBM-Levels-of-Evidence-2.1.pdf
[9] The Cochrane Collaboration. Cochrane handbook for systematic reviews of interventions. http://handbook.cochrane.org
[10] Lendvay TS, Casale P, Sweet R. Initial validation of a virtual-reality robotic simulator. J Robotic Surg 2008;2:145–9.
[11] Kenney PA, Wszolek MF, Gould JJ, et al. Face, content, and construct validity of dV-trainer, a novel virtual reality simulator for robotic surgery. Urology 2009;73:1288–92.
[12] Sethi AS, Peine WJ, Mohammadi Y, et al. Validation of a novel virtual reality robotic simulator. J Endourol 2009;23:503–8.
[13] Korets R, Mues AC, Graversen JA, et al. Validating the use of the Mimic dV-trainer for robotic surgery skill acquisition among urology residents. Urology 2011;78:1326–30.
[14] Hung AJ, Zehnder P, Patil MB, et al. Face, content and construct validity of a novel robotic surgery simulator. Urology 2011;186:1019–24.
[15] Lee JY, Mucksavage P, Kerbl DC, et al. Validation study of a virtual reality robotic simulator—role as an assessment tool? J Urol 2012;187:998–1002.
[16] Liss MA, Abdelshehid C, Quach S, et al. 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 2012;26:1629–34.
[17] Perrenot C, Perez M, Tran N, et al. The virtual reality simulator dV-Trainer® is a valid assessment tool for robotic surgical skills. Surg Endosc 2012;26:2587–93.
[18] Egi H, Hattori M, Tokunaga M, et al. Face, content and concurrent validity of the Mimic® dV-Trainer for robot-assisted endoscopic surgery: a prospective study. Eur Surg Res 2013;50:292–300.
[19] Schreuder HW, Persson JE, Wolswijk RG, et al. Validation of a novel virtual reality simulator for robotic surgery. Sci World J 2014;2014:507076.
[20] Kang SG, Cho S, Kang SH, et al. The Tube 3 module designed for practicing vesicourethral anastomosis in a virtual reality robotic simulator: determination of face, content, and construct validity. Urology 2014;84:345–50.
[21] Hung AJ, Shah SH, Dalag L, et al. Development and validation of a novel robotic procedure-specific simulation platform: partial nephrectomy. J Urol 2015;194:520–6.
[22] Alzahrani T, Haddad R, Alkhayal A, et al. Validation of the da Vinci surgical skill simulator across three surgical disciplines: a pilot study. Can Urol Assoc J 2013;7:E520–9.
[23] Kelly DC, Margules AC, Kundavaram CR, et al. Face, content, and construct validation of the da Vinci skills simulator. Urology 2012;79:1068–72.
[24] Lyons C, Goldfarb D, Jones SL, et al. Which skills really matter? Proving face, content, and construct validity for a commercial robotic simulator. Surg Endosc 2013;27:2020–30.
[25] Ramos P, Montez J, Tripp A, et al. Face, content, construct and concurrent validity of dry laboratory exercises for robotic training using a global assessment tool. BJU Int 2014;113:836–42.
[26] Seixas-Mikelus SA, Kesavadas T, Srimathveeravalli G, et al. Face validation of a novel robotic surgical simulator. Urology 2010;76:357–60.
[27] Chowriappa A, Raza SJ, Fazili A, et al. Augmented-reality-based skills training for robot-assisted urethrovesical anastomosis: a multi-institutional randomised controlled trial. BJU Int 2015;115:336–45.
[28] van der Meijden OA, Broeders IA, Schijven MP. The SEP "robot": a valid virtual reality robotic simulator for the da Vinci surgical system? Surg Technol Int 2010;19:51–8.
[29] Gavazzi A, Bahsoun AN, Van Haute W, et al. Face, content and construct validity of a virtual reality simulator for robotic surgery (SEP robot). Ann R Coll Surg Engl 2011;93:152–6.
[30] Seixas-Mikelus SA, Stegemann AP, Kesavadas T, et al. Content validation of a novel robotic surgical simulator. BJU Int 2011;107:1130–5.
[31] Finnegan KT, Meraney AM, Staff I, et al. da Vinci skills simulator construct validation study: correlation of prior robotic experience with overall score and time score simulator performance. Urology 2012;80:330–5.
[32] Hung AJ, Jayaratna IS, Teruya K, et al. Comparative assessment of three standardized robotic surgery training methods. BJU Int 2013;112:864–71.
[33] Connolly M, Seligman J, Kastenmeier A, et al. Validation of a virtual reality-based robotic surgical skills curriculum. Surg Endosc 2014;28:1691–4.
[34] Chowriappa AJ, Shi Y, Raza SJ, et al. Development and validation of a composite scoring system for robot-assisted surgical training—the Robotic Skills Assessment Score. J Surg Res 2013;185:561–9.
[35] Raza SJ, Froghi S, Chowriappa A, et al. Construct validation of the key components of Fundamental Skills of Robotic Surgery (FSRS) curriculum—a multi-institution prospective study. J Surg Educ 2014;71:316–24.

[36] Balasundaram I, Aggarwal R, Darzi A. Short-phase training on a virtual reality simulator improves technical performance in tele-robotic surgery. Int J Med Robot 2008;4:139–45.

[37] Hung AJ, Patil MB, Zehnder P, et al. Concurrent and predictive validation of a novel robotic surgery simulator: a prospective, randomized study. J Urol 2012;187:630–7.

[38] Moglia A, Ferrari V, Morelli L, et al. Distribution of innate ability for surgery amongst medical students assessed by an advanced virtual reality surgical simulator. Surg Endosc 2014;28:1830–7.

[39] Lerner MA, Ayalew M, Peine WJ, et al. Does training on a virtual reality robotic simulator improve performance on the da Vinci surgical system? J Endourol 2010;24:467–72.

[40] Cho JS, Hahn KY, Kwak JM, et al. Virtual reality training improves da Vinci performance: a prospective trial. J Laparoendosc Adv Surg Tech A 2013;23:992–8.

[41] Whitehurst SV, Lockrow EG, Lendvay TS, et al. Comparison of two simulation systems to support robotic-assisted surgical training: a pilot study (swine model). J Minim Invasive Gynecol 2015;22:483–8.

[42] Vaccaro CM, Crisp CC, Fellner AN, et al. Robotic virtual reality simulation plus standard robotic orientation versus standard robotic orientation alone: a randomized controlled trial. Female Pelvic Med Reconstr Surg 2013;19:266–70.

[43] Culligan P, Gurshumov E, Lewis C, et al. Predictive validity of a training protocol using a robotic surgery simulator. Female Pelvic Med Reconstr Surg 2014;20:48–51.

[44] Kiely DJ, Gotlieb WH, Lau S, et al. A randomized controlled trial of a proficiency-based, virtual-reality robotic simulation curriculum to teach robotic suturing. Gynecol Oncol 2014;133:193.

[45] Stegemann AP, Ahmed K, Syed JR, et al. Fundamental skills of robotic surgery: a multi-institutional randomized controlled trial for validation of a simulation-based curriculum. Urology 2013;81:767–74.

[46] Gallagher AG, O'Sullivan GC. Fundamentals of surgical simulation. London, UK: Springer Verlag; 2011.

[47] Maan ZN, Maan IN, Darzi AW, et al. Systematic review of predictors of surgical performance. Br J Surg 2012;99:1610–21.

[48] Gupta V, Lantz AG, Alzharani T, et al. Baseline urologic surgical skills among medical students: Differentiating trainees. Can Urol Assoc J 2014;8:242–6.

[49] Stone RJ, McCloy RF. Virtual environment training systems for laparoscopic surgery; activities at the UK's Wolfson Centre for Minimally Invasive Therapy. J Med Virtual Reality 1996;1:42–51.

[50] Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 2002;236:458–63.

[51] Hyltander A, Liljegren E, Rhodin PH, et al. The transfer of basic skills learned in a laparoscopic simulator to the operating room. Surg Endosc 2002;16:1324–8.

[52] Ahlberg G, Enochsson L, Gallagher AG, et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. Am J Surg 2007;193:797–804.

[53] Baheti A, Seshadri S, Kumar A, et al. RoSS: virtual reality robotic surgical simulator for the da Vinci surgical system. In: Proceedings of the 2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. IEEE; 2008;479–80.

[54] Lendvay TS, Casale P, Sweet R, et al. VR robotic surgery: randomized blinded study of the dV-Trainer robotic simulator. Stud Health Technol Inf 2008;132:242–4.

[55] Lawson Health Research Institute. Randomized controlled evaluation of robotic cardiac surgery training modalities. https://clinicaltrials.gov/ct2/show/study/NCT02357056?term=da+vinci+simulator&rank=2

[56] Sperry SM, O'Malley Jr BW, Weinstein GS. The University of Pennsylvania curriculum for training otorhinolaryngology residents in transoral robotic surgery. J Otorhinolaryngol Relat Spec 2014;76:342–52.

[57] Foell K, Finelli A, Yasufuku K, et al. Robotic surgery basic skills training: evaluation of a pilot multidisciplinary simulation-based curriculum. Can Urol Assoc J 2013;7:430–4.

[58] Fisher RA, Dasgupta P, Mottrie A, et al. An over-view of robot assisted surgery curricula and the status of their validation. Int J Surg 2015;13:115–23.

[59] Ahmed K, Khan R, Mottrie A, et al. Development of a standardised training curriculum for robotic surgery: a consensus statement from an international multidisciplinary group of experts. BJU Int 2015;116:93–101.

[60] Volpe A, Ahmed K, Dasgupta P, et al. Pilot validation study of the European Association of Urology robotic training curriculum. Eur Urol 2015;68:292–9.

[61] Zhang N, Sumer BD. Transoral robotic surgery: simulation-based standardized training. JAMA Otolaryngol Head Neck Surg 2013;139:1111–7.

[62] Teishima J, Hattori M, Inoue S, et al. Retention of robot-assisted surgical skills in urological surgeons acquired using Mimic dV-Trainer. Can UrolAssoc J 2014;8:E493–7.

[63] Rehman S, Raza SJ, Stegemann AP, et al. Simulation-based robot-assisted surgical training: a health economic evaluation. Int J Surg 2013;11:841–6.

[64] Roscoe SN. Incremental transfer effectiveness. Hum Factors 1971;13:561–7.

[65] Fletcher JD, Orlansky J. Recent studies on the cost-effectiveness of military training in TTCP countries. IDA Paper P-1896. Alexandria, VA: Institute for Defense Analyses; 1986.