Appendix 1.

Study characteristics including quality scores

Study of a Quantitative Method	Study Aim (Subjects of Healthcare Education)	Design (Participants)	Outcome Measures	Summary of Results	Application/ Technologies (Training time)	Display System	MERSQI Score (18)	Overall Rating (7)
Abhari et al. (2015)	Evaluation of an HMD-based guidance system compared with three planning environments (Resection planning of brain tumour from images and head phantom)	Single-group posttest (Study 1 and 2) (10 novices/non-clinicians) Two-group non-randomized comparison (Study 3) (7 clinicians and 14 novices/non-clinicians)	Test: 1) Difference in points of entry 2) Deviation between angles of surgical path 3) Accuracy 4) Response time 5) Index of performance	AR/MR significantly improved non-clinicians' performance (p<.01) compared to conventional planning environments (Study 1 and 2) AR/MR guidance significantly reduced the time of the task performed by clinicians (p<.05) (Study 3)	Self-developed for HMD with tracker recognizing physical and virtual representations of a head phantom. Connected with a foot pedal to interact with the system and to toggle between AR and MR (Not reported)	AR/MR	11.5	4
Aebersold et al. (2018)	Preliminary evaluation of a procedure training application (Simulating nasogastric tube (NGT) insertion on phantom)	Mixed methods study: Randomized controlled trial (RCT) and survey (69 nursing students, Control=34; AR=35)	Test: 1) Self-developed checklist for performance Questionnaire: 2) Likert scale on LE	Statistically significant correct placement of NGT through all checklist items in the AR group vs. control (p<.011). Participants' agreed /strongly agreed that AR was better for visualization (p<.01) and useful as tool in skill training (p<.015)	Company-developed application for mobile devices (20-25 minutes)	AR	15.5	5

Albrecht, Folta-Schoofs, Behrends, & Von Jan (2013)	Comparative study of an application (Learning of gunshot wounds)	Mixed methods study: RCT (pretest and post-test) and survey (10 medical students, Control=4; AR=6)	Test (pre- and post-completion): 1) Self-developed single choice (improvement) Questionnaire: 2) AttrakDiff2 (Likert scale) on LE 3) POMS on Mood States (pre- and post-completion) Observation (by non-participants): Directly on learning behavior	The test score was significantly improved in AR group (p<.03) Hedonic quality was significantly favored by AR group (p<.005). Fatigue and numbness significantly decreased, and vigor rose in the AR group. Observations showed interactive discussion in AR group vs. individual approach in control group	Self-developed application for mobile devices recognizing markers overlaying images onto user's body (30 minutes)	AR	14.5	4
Bifulco et al. (2014)	Investigation of the feasibility of an HMD-based application (Recording an electrocardiogram (ECG) on phantom and healthy patient)	Two-group non-randomized comparison (20 non-clinicians, manikin=10; patient=10)	Test: 1) Accuracy (average errors in mm) 2) Displacement errors (max error)	Average positioning errors of precordial electrodes were better on phantom vs. healthy patient. Max errors for the V6-lead <16 mm in both tests did not exceed clinical threshold of 25 mm	Self-developed for HMD with webcam recognizing markers attached to ECG device and phantom-patient (Few minutes)	AR	10.5	3
Ferrer-Torregrosa, Torralba, Jimenez, García, & Barcia (2015)	Comparison of an application (Learning anatomy of the lower limb)	Mixed methods study: RCT and survey (211 students of anatomy, Control=134; AR=77)	Test: 1) Self-developed multiple choice Questionnaire: 2) Self-developed on LE (metacognitive perception)	The AR group achieved significant better test result (p=.0001), and significantly surpassed the control group in terms of metacognitive perception (p<.05)	Self-developed for computer with webcam recognizing markers in printed book (Not reported)	AR	15.5	4

Ferrer-Torregrosa et al. (2016)	Comparison of a didactic aid based on AR with images and video (Learning anatomy of the foot muscles)	Mixed methods study: Three-group RCT and survey (171 students of anatomy, images/ Control= 60; Video=51; AR=60)	Test: 1) Self-developed Questionnaire: 2) Self-developed on LE (metacognitive perception) 3) Follow-up interview on learning success	Significant higher test score was obtained with aid of AR compared with video and notes (p<.000). The metacognitive perception was significantly favored by the AR group (p<.05), also sharing higher expectations for AR-based learning success.	Company-developed for mobile devices recognizing markers in printed book (14 days)	AR	13.5	4
Huang et al. (2018)	Investigation of the feasibility of an HMD-based application (Simulating US-guided CVC on phantom)	Mixed methods study: Prospective RCT and survey (32 novice operators, Control=16; AR=16)	Test: 1) Cannulation time 2) Procedure time 3) Adherence level Questionnaire: 4) Expert-developed on LE (usability and ergonomics)	No significant difference in cannulation time (p=.09) or procedure time (p=.29) for the AR group vs. Control. Adherence level were significantly favored by the AR group (p=.003). The majority >80% accepted the device in terms of ergonomics.	Self-developed for HMD rendering an instructional slide show connected to a computer and a foot pedal to navigate between the content (5-10 minutes)	AR	13.5	5
Jeon, Choi, & Kim (2014)	Investigation of a novel visualization device (Simulating US-guided CVC on phantom)	Prospective cross-over trial (20 physicians, Control/AR=20)	Test: 1) Time 2) No. needle redirections	Median of procedure time was clinically significant reduced by 50% in AR group vs. Control (p<.001). The number of needle-redirections significantly decreased in the AR group (p<.001)	Self-developed for micro projector attached to an ultrasound probe projecting images directly onto phantom (10 minutes)	AR	11.5	2

Keri et al. (2015)	Evaluation of a needle guidance system (Simulating lumbar puncture on phantom with abnormal spine)	RCT (24 residents, Control=12; MR=12)	Test (without assistive MR): 1) Needle path 2) Tissue damage 3) Procedure time 4) Needle insertion time 5) Success rate	Residents trained with MR visualization had better performance metrics: The MR group outperformed the control group significantly for needle path (p=.02), tissue damage (p=.01) and needle insertion time (p=.05) but not procedure time (p=.06) or success rate (p=.99)	Company-developed for computer, ultrasound machine, and tracker sensor-recognizing a virtual model of a vertebral column registered to a physical phantom (20 minutes)	MR	12.5	5
Kugelmann et al. (2018)	Evaluation of the feasibility of a tutorial (Learning of human gross anatomy)	Prospective large-scale cross-over survey (880 medical students, Control/AR=880 /748 in survey)	Questionnaire: 1) Likert scale on LE 2) Advantages and disadvantages 3) 4-item rating of the tutorial	The students agreed that the system increased the motivation 59% and greatly improved 3D understanding 93.4% (strongly agreed). AR was found advantageous to traditional books and rated ‘good’ by 81.9%	Company-developed for a computer connected to two cameras recognizing sensor-landmarks and overlaying images onto user’s body (Before/during the tutorial)	AR	7	2
Küçük, Kapakin, & Göktaş (2016)	Determination of learning effect via mobile AR (Learning of neuroanatomical pathways)	Mixed methods study: RCT and survey (70 medical students, Control=36; AR=34)	Test: 1) Self-developed multiple choice 2) Self-translated Cognitive Load (Likert) Scale Questionnaire: 3) Interview on LE	Achievement was significantly higher (p<.05) and cognitive load significantly lower reported in AR group (p<.05). Of students in AR group 79% responded that mobile AR facilitated learning the subject	Company-developed for mobile devices recognizing markers in printed book (5 hour-course)	AR	14.5	5

Leitritz et al. (2014)	Evaluation of the usability of an HMD-based application for examination (Training ophthalmoscopy on head phantom and test person)	Mixed methods study: RCT and survey (37 medical students, Control=18; AR=19)	Test: 1) Accuracy (No. of sketched vessels) 2) Self-developed (OTS) score Questionnaire: 3) Likert scale on LE (self-evaluation)	Significantly higher accuracy (p<.0083) and OTS vs. Control (p<.0033), but self-evaluation was not significantly different between the two groups	Company-developed for HMD connected to computer recognizing a model lens and a head phantom (15 minutes)	AR	14.5	4
Ma et al. (2016)	Investigation of precision of a personalized system (Learning of human gross anatomy)	Two single-group post-tests and survey (Study 1) (2 surgeons and 5 medical students) (Study 2) (72 medical students)	Test (quantified by participants): 1) Accuracy (Study 1) Questionnaire: 2) Likert scale on usability 3) Likert scale on LE (Study 2)	Accuracy was demonstrated, and study participants favored the usability. The learning potential of AR was accepted by 86.1%, and found valuable as a display system of anatomy 91.7%	Company-developed for computer connected to two cameras recognizing sensor-landmarks and overlaying images onto user's body (15 minutes)	AR	7.5	2
Mewes et al. (2019)	Provision and evaluation of a needle guidance system (Simulating MR-guided needle insertion into calibration phantom)	Single-group posttest and survey (4 radiologists and 4 technicians)	Test: 1) Entry point error 2) Target point error 3) Insertion time Questionnaire: Expert-interview on LE (usability)	The targets were reached, and the answers of the users were predominantly positive supporting the suitability of the system	Self-developed for projector coupled to two cameras inside a wide-bore MRI scanner recognizing markers on phantom (Until users felt confident)	AR	10.5	3
Moro, Štromberga, Raikos, & Stirling (2017)	Comparison of an AR module with two learning modes (virtual reality (VR) and tablet) (Learning of skull anatomy)	Mixed methods study: Three-group RCT and survey (59 health science students, tablet/Control=22; VR=20; AR=17)	Test: 1) Self-developed multiple choice Questionnaire: 2) Scale on adverse health effects 3) Likert scale on LE	No significant difference in test scores between the three learning modes (p<.874). Adverse effects as dizziness were significantly experienced in the VR group vs. AR and tablet group (p<.001). Perception of AR was high but not significant	Self-developed for mobile devices (10 minutes)	AR	13.5	5

Moult et al. (2013)	Evaluation of a needle guidance system (Simulating diagnostic US-guided facet joint injections on phantom)	RCT (26 pre-medical undergraduate students, Control=13; MR=13)	Test (without assistive technology): 1) Success rate 2)Total time 3) Time inside 4) Total path 5) Path inside	Significantly higher mean success rate of 61.5% in MR group vs. Control 38.5% (p=.031). No significant difference was found in any of the needle metrics of procedure times or path lengths	Company-developed for computer, ultrasound machine, and tracker sensor-recognizing a virtual model of a vertebral column registered to a physical phantom. (10 minutes)	MR	13.5	4
Noll, Von Jan, Raap, Albrecht, & Albrecht (2017)	Comparison of an AR application with mobile blended learning environment (Diagnosing various skin diseases)	Mixed methods study: RCT (pretest, posttest, follow-up) and survey (44 medical students, mobile phone/Control=22; AR=22)	Test (pre-, post- and follow-up-completion): 1) Self-developed single choice (improvement) 2) Retention (average decrease of correct answers) Questionnaire: 3) AttrakDiff2 on LE 4) POMS on Mood States (pre- and post-completion)	No significant difference in test score or retention of knowledge. No significant variations were found regarding experience and emotions between the groups of AR and mobile blended learning	Self-developed application for mobile devices recognizing markers overlaying images onto user's body (45 minutes)	AR	14.5	6
Rai, Rai, Mavrikakis, & Lam (2017)	Validation and assessment of the efficacy of an HMD-based application (Training ophthalmoscopy on head phantom)	Prospective three-group RCT (28 novice residents and 3 fellows (experts), Control=15; AR=13; No training=3 (experts))	Test: 1) Total time 2) Total score 3) Performance (task scores/time)	Time required was not significantly different (p=.11), but the AR group significantly demonstrated superiority in total score (p=.02) and performance (p=.006). Fellows outperformed novice residents despite no prior experience with simulator	Company-developed for HMD connected to computer recognizing a model lens and a head phantom (About 2 hours)	AR	14.5	5
Robinson et al. (2014)	Evaluation of a new MR part-task trainer	Mixed methods study: Three-group non-randomized comparison and survey	Test (pre- and post-intervention without assistive technology): 1) SCVA score	All participants significantly improved SCVA score (p<.0001) and time (p<.0001). The participants	Self-developed for computer with tracker sensor-recognizing a virtual model of the phantom registered within a	MR	13.5	7

	(Simulating subclavian venous access (SCVA/CVC) without US-guidance on phantom)	(65 physicians of different training categories, novices=25; intermediates=24; experts=16)	2) Time 3) No. attempts 4) No. skin punctures 5) Success rate 6) Complication rates (pneumothorax and subclavian puncture) Questionnaire: 5) Likert scale on LE (usability) 6) Likert scale on performance confidence (pre- and post-intervention)	significantly reduced no. attempts (p<.0001), no. skin punctures (p=.0007), but no significant difference was found though success rate was increased (p=.08). Both complication rates fell with MR. The majority 95.4% strongly agreed the usability for future CVC. Confidence significantly rose (p<.0001)	3D-printed phantom built-up of head and thorax CT scan (Until users felt confident)			
Rochlen, Levine, & Tait (2017)	Evaluation of usability of an HMD-based needle guidance system (Simulating CVC without US-guidance on phantom)	Mixed methods study: Two-group non-randomized comparison and survey (40 medical students /participants, No prior CVC training=13; prior CVC training=27)	Test: 1) Correct identification 2) Correct needle insertion (accuracy) 3) Time Questionnaire: 4) Likert scale on LE 5) Open-ended evaluation (ergonomics)	No significant difference in identification, needle insertion, and time expense between experienced and non-experienced. Participants favored AR in visualizing anatomy 92.5% and for incorporation into training 82.1%. Evaluation addressed issues of poor ergonomics <44.4%	Self-developed for HMD with external camera recognizing markers on needle and phantom (Until users felt confident)	AR	14	3
Siebert et al. (2017)	Comparative investigation of adherence to a guideline adapted for HMD (Simulating pediatric cardiopulmonary resuscitation on phantom)	Mixed methods study: Prospective RCT and survey (20 residents, pocket reference cards/Control=10; AR=10)	Test (deviation from guidelines): 1) Time to first defibrillation/DF 2) Time to first compression 3) Drug and shock doses 4) No. of shocks Questionnaire: 5) Likert scale on LE (stress perception)	Adherence by time to first DF and compressions were not improved, but errors were significantly reduced in administering shock doses vs. Control (p<.001). No significant difference in stress response (p=.38)	Self-developed for HMD rendering guideline cards in the glasses with touchpad to navigate between the content (15 minutes)	AR	13.5	6

Solbiati et al. (2018)	Preliminary assessment of a needle guidance system (Simulation CT scan-guided needle insertion into phantom, porcine, and cadaver)	Single group posttest (proof-of-concept study) (Study participants not specified)	Test: 1) Computed accuracy (mm)	An acknowledged targeting accuracy was achieved in all cases but in the breathing porcine model	Self-developed for mobile devices recognizing markers on tool and phantom-porcine-cadaver. (Not reported)	AR	8.5	2
Sutherland, Hashtrudi-Zaad, Sellens, Abolmaesumi, & Mousavi (2013)	Demonstration of the potential and functionality of an application (Simulating US-guided spinal needle insertion on phantom)	Two-group non-randomized comparative survey (10 participants, residents=4; students and technicians=6)	Test: 1) Force (traversing of tissue) Questionnaire: 1) Likert scale on LE (functionality)	Peak values of the forces and the pattern of the profile corresponded to related work. The system was positively reviewed on the system regarding functionality, visual feedback, and haptic feedback	Self-developed for computer coupled to a haptic device with stylus and camera recognizing sensors attached to a dummy ultrasound probe and a phantom. (5-10 minutes)	AR	9.5	2
L. L. Wang, Wu, Bilici, & Tenney-Soeiro (2016)	Implementation and demonstration of a prototype (Test preparation for neurologic clinical shelf exam)	Single-group survey (24 medical students)	Questionnaire: 1) Query of LE (utility)	Upon demonstration 100% of participants agreed that AR improved the learning capacity for the textbook	Self-developed for mobile devices recognizing markers in printed book (Demonstration)	AR	7	1
Wang et al. (2017)	Evaluation of feasibility and user experience of an HMD-based telemedicine mentoring platform (Training US examination for trauma on healthy patient under guidance of mentor)	Three-group non-randomized comparison and survey (24 medical students and 1 mentor, Full telemedicine setup/Control=12; AR=12; mentor=1)	Test: 1) Expert-Global Rating Scale for performance 2) Completion time Questionnaire: 3) Likert scale on LE (utility) 4) Cognitive load	Performance of the AR group was not significantly improved (p=.534), but the AR group had a significant prolonged completion time (p=.008). The AR group showed no significant difference though they favored the utility of AR (p=.065) and reported a lower cognitive load (p=.28)	Self-developed for HMD with an ultrasound probe connected to computer and live-streamed to mentor connected to a sensor-controller projecting mentor's hands and gestures back into the AR space of the trainees (No prior training)	AR	12	7

Zhu, Fors, & Smedberg (2018)	Exploration of needs and challenges in applying AR in continuing professional development (CPD) (Training of general practitioners within primary care in China)	Qualitative semi-structured face-to-face interviews (13 physicians and 2 managers)	Questionnaire: 1) Interview on attitudes toward usage 2) Query of suitability for subjects in future	The participants reacted positively to usage of AR in CPD, especially concerning visualization and skill training. The design should improve competencies, understand learning needs, and stimulate positive attitudes toward technology	Company-developed application for mobile devices (Demonstration)	AR	12 (AQRAME) (12)	6

KEY: HMD, head-mounted display; AR, augmented reality; MR, mixed reality; LE, learning experience; CVC, central venous catheterization; US, ultrasound