International Journal of Medical Education. 2020;11:1-18
ISSN: 2042-6372
DOI: 10.5116/ijme.5e01.eb1a

Augmented reality and mixed reality for healthcare education beyond surgery: an integrative review

Jaris Gerup[1], Camilla B. Soerensen[2], Peter Dieckmann[3]

[1]School of Medical Sciences, University of Copenhagen, Denmark
[2]Department of Pediatrics, Herlev and Gentofte Hospital, Denmark
[3]Copenhagen Academy of Medical Education and Simulation (CAMES), Center for Human Resources, Herlev and Gentofte Hospital, Denmark

Correspondence: Jaris Gerup, Copenhagen Academy of Medical Education and Simulation (CAMES), Herlev and Gentofte Hospital, Herlev Ringvej 75, 25th Floor, 2730 Herlev, Denmark. Email: jaris.gerup@gmail.com

Accepted: December 24, 2019

Abstract

Objective: This study aimed to review and synthesize the current research and state of augmented reality (AR), mixed reality (MR) and the applications developed for healthcare education beyond surgery.

Methods: An integrative review was conducted on all relevant material, drawing on different data sources, including the databases of PubMed, PsycINFO, and ERIC from January 2013 till September 2018. Inductive content analysis and qualitative synthesis were performed. Additionally, the quality of the studies was assessed with different structured tools.

Results: Twenty-six studies were included. Studies based on both AR and MR involved established applications in 27% of all cases (n=6), the rest being prototypes. The most frequently studied subjects were related to anatomy and anesthesia (n=13). All studies showed several healthcare educational benefits of AR and MR, significantly outperforming traditional learning approaches in 11 studies examining various outcomes. Studies had a low-to-medium quality overall with a MERSQI mean of 12.26 (SD=2.63), while the single qualitative study had high quality.

Conclusion: This review suggests the progress of learning approaches based on AR and MR for various medical subjects while moving the research base away from feasibility studies on prototypes. Yet, lacking validity of study conclusions, heterogeneity of research designs and widely varied reporting challenges transferability of the findings in the studies included in the review. Future studies should examine suitable research designs and instructional objectives achievable by AR and MR-based applications to strengthen the evidence base, making it relevant for medical educators and institutions to apply the technologies.

Keywords: Augmented reality, mixed reality, healthcare education, medicine, integrative review

Introduction

The integration of digital strategies has brought healthcare education to a paradigm shift, now reflected in many educational curricula.[1] Modern teaching curricula aim to educate trainees efficiently in safe environments to establish transferability into the clinical context. Augmented reality (AR) and mixed reality (MR) have long been expected to be disruptive technologies, with potential uses in medical education, training, surgical planning and to guide complex procedures.[2] While virtual reality (VR) has mainly led the way for the implementation of the display technologies, it is criticized for several limitations.[3,4] The term display technologies will hereafter be used to refer to AR and MR although it in principle also covers VR. The latter, however, is beyond the scope of this review.

AR describes display-based systems that combine real and virtual imagery, which are interactive in real-time and register the real-world environment to be augmented by virtual imagery.[5] The visual display technology augments the physical environment by especially two principal manifestations: See-through (transparent) head-mounted display and non-immersive monitor-based video (window on the world).[6] AR systems are based on the combination of the physical

and the virtual environment. On the contrary, in VR systems the participant is totally immersed in a completely virtual one.

MR is defined as the merging of real and virtual worlds and can be seen as a larger class of technologies covering the display environment of AR and augmented virtuality (AV).[7] Where virtual information augments the real view in AR, real-world information augments the virtual scene in AV. The external inputs providing real-world context are also seen in VR but were classified as MR in this review. The term of MR was included to embrace new technology labeled as MR, that tries to define a clear distinction between AR and MR, even if there is none.[8]

The abilities to provide situated and authentic experience connected with the real environment, enhance interaction between the physical and virtual content, while preserving a feeling of presence explains the growing expectations that AR and MR may be suitable for healthcare education in various contexts.[9]

Concerning healthcare education, the process of teaching, learning and training with an ongoing integration of knowledge, experience, skills and responsibility qualifies an individual to practice medicine.[10] Looking into medical education, several authors request to eliminate outdated, inefficient, and passive learning approaches and start to embrace these newer methodologies of learning.[11] Surgeons have historically always been quick to adapt to new technology developing new treatment and learning methodologies, while physicians were rather more tardy.[12] Today most studies on display technologies stem from surgery. In an integrative review on AR in healthcare education from 2014, surgical studies accounted for 64% (n=16) of the studies included.[13] A recent systematic review on AR for the surgeon clarifies the current lack of systematic reviews for physicians and ultimately educators within the field of medicine.[14] Many internists and other medical specialists do no longer diagnose and treat illnesses using only their knowledge of pathophysiology and pharmacology.[15] Today, many physicians have taken up procedures and surgical treatment initiatives by operation or manipulation defined as the use of hands to produce the desired movement or therapeutic effect in part of the body.[16] Nevertheless, medicine consists essentially of non-surgical treatment, procedures and other approaches of diagnostics and prevention of disease that need to be taught, learned and trained with an ongoing evaluation of adaptations. AR and MR may effectively help medical educators achieve such instructional objectives for medical education as it is being used for surgical training.

According to the review by Zhu and colleagues, publications in the field of AR increased significantly in 2008.[13] Now, ten years after that publication outbreak, a new review is warranted. To the best of our knowledge, current reviews on AR and MR have not specifically studied applications for medical subjects in healthcare education. Most papers predominantly include surgical studies and only a few focused on AR in either otolaryngology or medical training.[1,3,4,9,13,17] Currently, no adequate reviews are available that uncover the educational profile of both AR and MR-based applications across different medical specialties, subjects and target groups.

Our aim of this integrative review was to investigate the current research and state of AR and MR-based applications for healthcare education beyond surgery, providing an overview of the findings, strengths and weaknesses of the reported studies.

Methods

We chose to conduct an integrative review, given that previous reviews showed only a few studies relevant for the current scope.[3,4,13,17] This is thought to be the broadest type of review as it allows the inclusion of various research designs and information sources.[18] The method also integrates a process of quality assessment of the studies included that may qualify the integrative review for recommending practice and answering complex search questions.[19,20] The digital databases of PubMed, PsycINFO and ERIC were searched. The journal of Medical Teacher was hand-searched. Ted Talks and podcasts on the iTunes Podcast app were included, acknowledging the increasing importance of "new media".[21,22] Studies published between January 2013 and September 2018 were included. Relevant word groups, combinations and open-ended terms used for the search were: "Augmented reality OR mixed reality" AND "medicine OR medical OR healthcare" AND "educat* OR simulat* OR train* OR learn*". We did not implement any filter of 'NOT virtual reality OR surgery' in our search string to avoid missing relevant studies examining non-surgical elements despite being termed as a surgical study.

Eligibility criteria

The selection process was done according to three overall criteria regarding research, focus on technology and content. According to the criterion of research studies were included if they described 1) a goal or research question, 2) an appropriate study design, 3) data collection and analysis methods and 4) the discussion of results. Research articles were excluded if they 1) neither described goal nor research question, 2) were review papers and 3) were focused on system descriptions without evaluation or other data. Table 1 provides the inclusion and exclusion criteria for the study.

Study selection

All abstracts were read by JG, who assessed whether they met the inclusion criteria. In case of doubt, JG discussed the inclusion of studies with the other authors. All duplicates were removed.

Data extraction and synthesis

Study characteristics and information of all articles were extracted and described by JG. Characteristics were authors, study aim, subject of healthcare education, design,

participants, outcome measures, results, application/technologies, training time and display system. Content analysis was used to describe the study designs and to inductively identify the strengths and weaknesses of AR and MR as described by the studies included.

Table 1. Inclusion and exclusion

Criterion	Inclusion criteria	Exclusion criteria
Research	• Goal or research question described • Study design described and appropriate • Data collection and analysis methods were described • Results were described and discussed	• Neither goal nor research question described • Review papers • System description without data evaluation
Focus on technology	• Combination of real and virtual environments • Interactive in real-time • Real or perceived registration in 2D or 3D	• Used augmented or mixed reality in name but investigated only virtual reality
Content	• Healthcare education • Medical education	• Education without medicine or only surgical focus • Medicine without education or only treatment or rehabilitation focus • Patient education related to treatment • Dentistry, veterinary medicine or other fields of education

Quality assessment

The methodological quality of quantitative and mixed methods studies was evaluated with the Medical Education Research Study Quality Instrument (MERSQI).[23] This 10-item instrument has been thoroughly assessed and evaluated for its correlation with other assessment tools for research quality.[24] MERSQI covers six domains of studies: Study design, sampling, type of data, the validity of evaluation instrument, data analysis and outcome. All domains assign 0-3 points valuing the study to a final score between 0 and 18, the larger number indicating better study quality. The score will be presented as mean, standard deviation (SD) and range in parentheses. Each study was scored at the highest possible level. If a study reported more than one outcome, the rating for the highest outcome score was recorded not differentiating between primary or secondary outcome.

The quality assessment of all studies was done by JG. In addition, to assess the quality of JG evaluation, a level of approximately 20% of the studies were randomly selected for assessment by co-authors and independently evaluated by at least two authors. We computed the intraclass correlation coefficient (ICC) to calculate the inter-rater reliability (IRR) between all authors.

The methodological quality of qualitative studies was evaluated with a 12-item grid for Appraising Qualitative Research Articles in Medical Education that was converted into a quality assessment tool (AQRAME) by the authors of this review.[25] The instrument covers five domains: Introduction, methods, results, discussion and conclusion. The domain of methods assigns 0-5 points and the conclusion domain only assigns 0-1 point, while the three remaining domains assign 0-2 points. It includes a score range between 0 and 12 points, with a larger number indicating better study quality. A score of 0.5 was given in case of an unclear answer of neither yes nor no. The score will be presented as mean, SD and range in parentheses.

An overall quality assessment tool was developed for rating all included studies regardless of their methodological design, assigning a figure of 1 to 7, with the larger number indicating better study quality. This was introduced to challenge the relative judgements of the MERSQI and AQRAME, acknowledging that different research questions inherently require different study designs. The appraisal was based on the need to be explicit about the role and assessment of the researcher in qualitative research.[26] For studies with mixed-method designs, we applied the MERSQI tool only, rating the quantitative parts of the study.

Results

Out of the 315 papers initially identified, four duplicates were removed, three articles in Chinese excluded, and one article could not be retrieved. No reporting of research was found in 14 Ted Talks and iTunes podcasts. Three hundred seven publications were screened and 281 excluded as they did not meet the inclusion criteria. Study subjects related to nasogastric tube insertion, facet joint injection, catheterization or needle guidance were interpreted to clinically related to medicine as a practice of diagnosis and so these studies were classified to fulfill the inclusion criteria. One study focusing on resection planning was included and categorized as preoperative visualization.[27] However, needle insertion itself was interpreted not to produce a desired movement or therapeutic effect in part of the body and not classified as a surgical procedure. This resulted in a total of 26 studies being included in the integrative review. The flow chart of publications selected for inclusion in this integrative review is displayed in Figure 1.

Study characteristics

The studies applied AR and MR primarily by integrating the display technologies into knowledge platforms and guidance systems for simulator practice. Some studies offered feedback in the endeavor of a skill or a field of knowledge, while others provided an immersion into scenarios and remote assessment-training for telemedicine. The display technologies showed the ability to stimulate the learning process and support the learner for several competencies: To understand

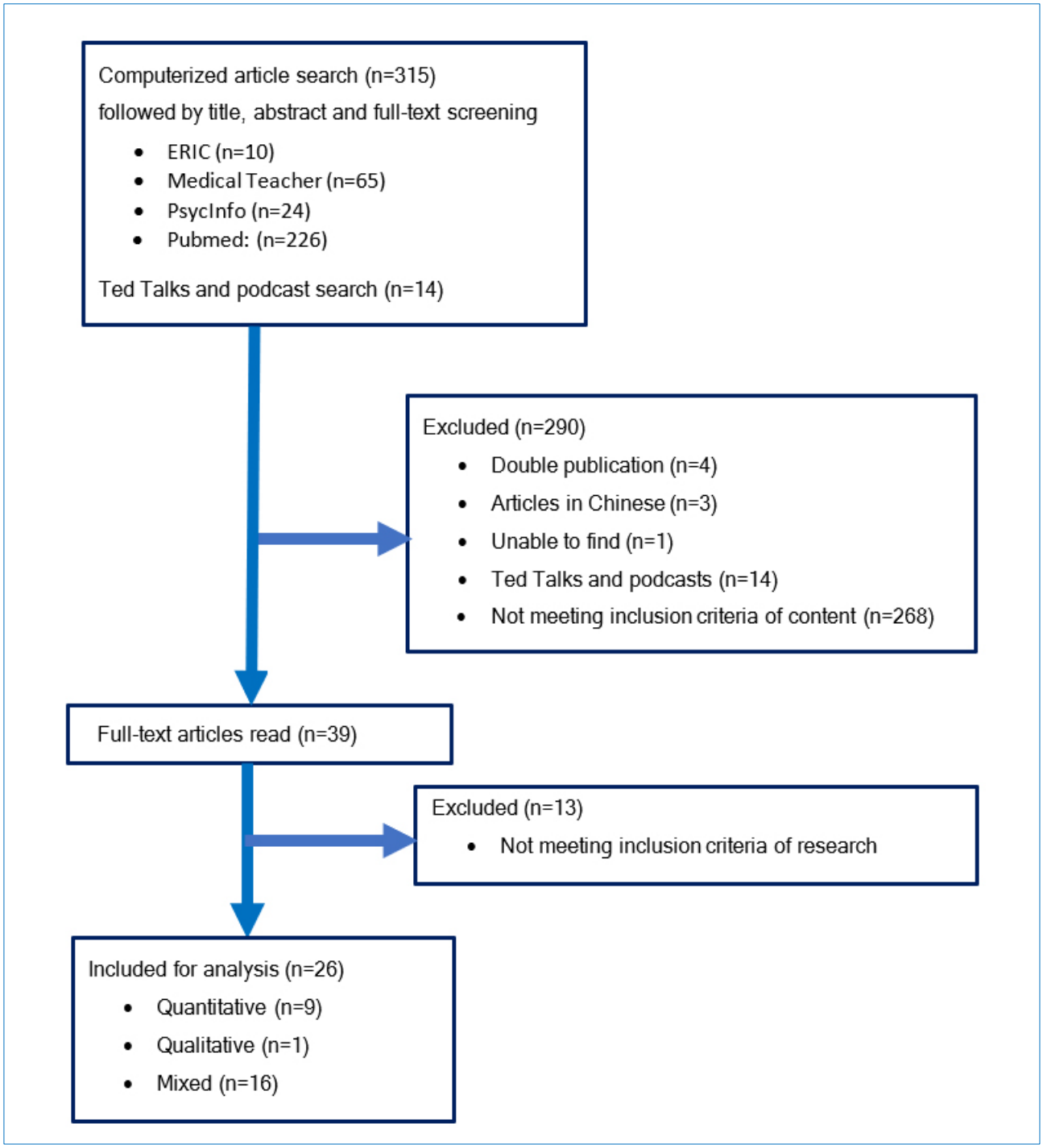

Figure 1. Selection process of studies

spatial relationships and construct mental 3D models of anatomy with the help or without 2D imaging. To acquire cognitive-psychomotor abilities, prolong learning retention, experience student-centered motivation and obtain flexibility to learn anytime and anywhere in their own pace and style. Furthermore, the studies suggested that AR and MR could complement practice in safe simulation environments contributing to patient safety and a higher degree of confidence (See Appendix 1 – "Summary of results").

Technical specifications

The majority of studies (n=22) examined an actual application of AR.[28–49] The rest (n=4) investigated an application based on MR.[27,50–52] Six applications developed by companies were reported in 10 studies.[30,31,37,39,40,43,47,48,50,51] The remaining studies (n=16) involved self-developed applications primarily developed at universities and hospitals.

Mobile device-based (tablets and smartphones) applications were used in nine studies.[33,35,37,39,41,42,47-49] Of these two

thirds (n=6) involved camera and marker-based recognition, and three studies did not report any further on the applications developed.[41,47,48] Eight studies implemented head-mounted display.[27,28,38,40,43-46] Two studies utilized the same head-mounted display.[40,43] The head-mounted display-integrated applications had marker-based recognition in four of the studies.[28,40,43,44] One study recognized the hands and gestures of a mentor projecting these into in the trainee's display.[46] Two studies implemented a foot pedal to interact with the application.[27,38] For one study this included toggling between AR and MR-mode.[27] Computers were used in 11 studies.[30,31,34,36,38,40,43,46,50–52]These delivered the computing power for head-mounted display-based applications in four studies.[38,40,43,46] One computer-based application had marker-based recognition.[36] Seven studies were sensor-based.[30,31,34,46,50–52]Two studies recognized landmarks of the user's body.[30,31] Four studies recognized a virtual model registered with a phantom characterized as MR.[27,50–52] Eleven studies reported using external cameras and tracking devices.[27,28,31,32,34,36,44,50–52]Two studies used applications based on projectors, one recognizing markers on a phantom, and one projecting images directly onto a phantom without using a tracking device.[29,51]

Methodological quality

In the included 26 studies, nine were solely quantitative, 16 were mixed research methods and one was qualitative. Based on rating comparisons of the approximately 20% (n=5) randomly selected papers, the authors' agreed to use the ratings by JG for MERSQI, AQRAME and the overall score for the remaining papers. The average total MERSQI score of the 25 quantitative and mixed methods studies was mean 12.26, SD=2.63 (7-15.5). The ICC between all raters were computed to IRR=.50 for the MERSQI overall score, which corresponds to a moderate reliability.[53] Nearly one-third of all studies (n=8) either had no evaluation tool or did not report any validity of the instrument used.[28–35]

The qualitative study involved semi-structured face-to-face interviews that explored the needs and challenges of applying AR for healthcare education. The study demonstrated a detailed clarity and rigor according to the individual AQRAME score of all three authors corresponding to 12 (JG), 11.5 (CBS), and 12 (PD). As there was only one qualitative study, we did not report any IRR for the AQRAME overall score.

The mean average overall quality score of all studies was 4.08, SD=1.65 (1-7) with an adjusted ICC equaling IRR=.429 also corresponding to a moderate reliability.[53] The scores of the individual studies and the study characteristics are reported in Appendix 1.

Strengths and weaknesses of AR and MR

Three themes were inductively identified indicating the strengths and weaknesses of AR and MR in healthcare education beyond surgery.

Strengths

Implemented across various subjects for learner types of all levels spanning different sectors

The most frequently studied subjects of healthcare education were found within anatomy (n=6) and anesthesia (n=7), the ladder represented by four studies focusing on central vein catheterization.[29,38,44,52] Study participants were divided into 12 different categories: Pre-medical, medical, nursing, and health science students, novices, residents, fellows and established clinicians of different specialties, technicians, non-clinicians, non-specified participants and managers. The mean number of participants was 77.1, SD=170.6 (1-880) since the sample size was set to one in a study that did not report or specify the study participants.[33] The distribution of studies across subjects of healthcare education related to the number of participants enrolled is described in Appendix 2.

The rich diversity of research and outcome focus

A total of six proof-of-concept, pilot or user studies sought to introduce an application or assess initial validity.[28,29,33–35,47] Eight studies focused on evaluating training by an application for strengthening the validity of the construct.[30,37,39,40,42,43,50,51] The remaining studies (n=12) focused on the application-based assessment of a specific skill or procedure, eventually correlating the performance to other outcomes such as cognitive load.[27,31,36,38,41,44-46,48,49,51,52] Technical test outcomes were reported in 17 studies and concerned primarily needle insertion in terms of accuracy and precision (n=11).[27-29,31,33,40,43,44,50-52] The secondly most reported technical test outcome concerned procedure time (n=9).[27,29,38,43,44,46,50-52] Nineteen studies investigated learning experience and user acceptance based on especially Likert scales.[30-32,34-42,44-49,52] Other questionnaire-based outcomes were cognitive load, stress response, adverse health effects and ergonomics.[38,39,41,44-46] Knowledge tests were examined in combination with questionnaire-based outcomes in six studies.[36,37,39,41,42,49] One study included an observational method to determine learning behavior.[49]

Growing evidence for improving learning

In 11 studies AR and MR were claimed to significantly improve the learning process or part-tasks associated in all or in the majority of outcome measures.[27,29,36,37,39,40,43,48-50,52]Four out of six studies examining the acquisition of anatomy knowledge reported significantly improved learning.[36,37,39,49] Significant positive findings were found in six of 11 studies concerning skill training of needle insertion favoring both students and established clinicians.[27,29,40,43,50,52] Procedure time was significantly reduced in three of nine studies.[27,29,52] Examining different questionnaire-based aspects of the learning experience and user acceptance four of 19 studies demonstrated significant positive findings advocating the usability

of the display technologies.[36,37,39,48] Fifteen studies found no significant positive results but all suggested the AR and MR-based applications may outperform traditional learning approaches within the involved subjects of healthcare education.[28,30-35,38,41,42,44-47,51] Other promising learning factors facilitated by the display technologies were related to visualization, directing attention, intrinsic benefits of motivation, physical interaction activating kinesthetic schemes, patient safety, skill retention, simulation confidence related to transferability, mobile learning and using oneself as a learning object.[39,41,42,45,49,51]

Weaknesses

Reporting of prototypes, technological limitations and poor ergonomics

Sixteen studies presented a prototype, typically as preliminary feasibility studies lacking to report adequately on the educational impact of the prototype tested.[27-29,32-36,38,41,42,44-46,49,52] Ten studies were conducted on one of six established applications.[30,31,37,39,40,43,47,48,50,51] The studies of head-mounted display-based applications (n=8) addressed technological limitations related to limited computing power, occlusion of the user's field of view and poor ergonomics by head-mounted displays being tethered to workstations and when wearing glasses underneath.[27,44,46]

Shortcomings of the study designs for transferability

Four studies were designed as a single group user study only, making strong conclusions difficult.[31-33,35] Twenty-two studies used a group design or comparison, of which the most (n=17) compared two groups.[27-30,34,36,38-40,42,44,45,47-51]Only two studies did not compare AR or MR with another media corresponding to lectures, books, video, virtual reality, mobile devices, conventional training platforms, and telemedical full-setup.[28,34] Two studies compared the media of mobile devices after having provided AR content to one of the groups.[41,42] Five studies encompassed three groups.[37,41,43,46,52] Two of the two-group studies used a cross-over design.[29,30] No study involved patients in an authentic context, but two studies included patient data.[27,32]

Lacking evidence for improving learning

Eight studies reported descriptive frequencies of self-reported evaluations and measures without any statistical analysis of significance.[28,30–35,47] Seven studies claimed the display technologies offered no significant impact for improving learning in all or in the majority of outcome measures.[38,41,42,44-46,51] The two studies that compared AR within the same media of mobile devices found no significant difference in any of the outcome measures.[41,42] Only a single study presented a significant negative finding of prolonged completion time of an ultrasound examination in the AR group.[46]Potentially conflicting factors were addressed in terms of visual misperception, media or technology enthusiasm-based motivation, negation of patient discomfort related to patient safety, and missing translation of performance from simulation to clinical setting.[27,41,50,51]

Discussion

Virtual augmentation and guidance of AR and MR are increasingly used in applications for medical subjects of healthcare education these years. The quality of the existing studies and applications including the educational benefits of the display technologies remain unclear at the moment.

We reviewed the current research and state of AR and MR-based applications for healthcare education in medical disciplines beyond surgery. Our integrative review identified 26 original studies examining various applications of both display technologies. The applications were found to measure numerous outcomes related to the learning process, acquisition of knowledge and skill training while providing feedback on patient care-related outcomes such as complication rates, insertion time and needle path related to tissue damage. This differs greatly from the findings of a systematic review by Barsom and colleagues on applications for medical training for professionals, in which none were developed to measure the prevention of errors for the interest of patient safety.[4]

Our work revealed an increased emergence of established applications corresponding to 27% (n=6) investigated in 10 studies against 16 prototypes. A prior review by Zhu and colleagues only found one established application for laparoscopic colorectal surgery.[13] In the same review, the authors found the application designs lacking guidance by learning theories only resting on traditional learning strategies. We observed that the applications of AR and MR still have not exploited the integration of learning theories and strategies into their design. Still, the increased number of established applications is a step towards turning the research base away from feasibility studies examining prototypes.

We conclude that the studies overall were of low-to-medium quality. This is consistent with the low to modest strength of evidence level reported in previous systematic reviews.[4,17] The single qualitative study was found to be of high quality in terms of clarity and rigor, while the relative judgement of the overall quality was found to be of a low-to-medium quality. The greatest limitation across the pool of studies noted in nearly one-third of all studies (n=8) was either the utter lack or poor reporting of the validity of the evaluation instruments indirectly providing the evidence base for the study findings. Additionally, the statistical analyses reported incomplete results or were unclearly interpreted. Shortcomings of the reviewed studies further included heterogeneity of research designs, unstandardized outcome measures and wide variation in details given. Widespread heterogeneity among studies is stated to be one of the greatest challenges of quantitatively synthesizing research evidence.[54] At the same time, an outspoken concern argues that media-comparative studies in learning are virtually useless and not valid for comparison.[55] From this perspective, the studies

failed to determine which media or technologies were best for healthcare education but rather informed practice with the specific application. These limitations are general for much education research but may be especially pronounced for research in the nexus of learning and technology.[56] Nevertheless, we did not exclude studies based on their quality due to our aim of providing an overview of the strengths and weaknesses of all relevant research in AR and MR for healthcare education beyond surgery during the past half-decade.

Limitations and recommendations for future studies

To our knowledge, this is the first integrative review of AR and MR solely focusing on medical subjects of healthcare education. Three articles in Chinese were not included, meaning that we possibly excluded relevant knowledge. Moreover, we may have missed relevant research either published or not published in technical journals as our main focus was on databases for healthcare and education. Our finding that all included studies suggested or reported significant positive findings should be interpreted with caution since publication bias cannot be excluded. We tried to minimize the drop-out of relevant material by including unpublished work from new online sources such as TED Talks and the podcast media of iTunes. There was a contentious issue of the designs and presentations of these varying too extensively without enhancing the quality and usefulness of the review. Our study abstained from addressing the educational profile of AV compared to AR both being encompassed by MR. This could not be done due to a low number of studies measuring AV-based learning, possibly related to the impaired technologic and conceptual understanding of MR across the research field and industry. The quality of the included studies was assessed with the MERSQI scale, which revealed inconsistencies across a few domains in the process of rating. This was mainly due to missing information in the reviewed studies as well as a lack of clarity in the MERSQI guidelines. Though moderate reliability was found between all raters in the MERSQI and the overall quality assessment tool, one could argue that the sample size of the rating corresponding to approximately 20% (n=5) of the studies either hinders or disallows reliable calculations beyond descriptive analysis. Finally, the self-developed assessment tool of AQRAME has not been validated for quality scoring qualitative research despite relying on a known 12-item grid for quality appraisal. This tool was introduced since we were not aware of any validated evaluation instruments for quality assessment of qualitative research in healthcare education.

A variety of applications for subjects of healthcare education beyond surgery have been developed, and their benefits were supported by this integrative review. We expect that more research will be done on the field as more institutions will explore and apply applications based on AR and MR in the future. Randomized controlled trials should continuously be organized for evaluating clinical performance and patient-care related outcomes. Specifically, the actual effects on real patients and physician behaviors towards patients in a real context are yet to be elucidated. We recommend future studies to justify and validate metrics and report the reliability of measures for higher-quality evaluations. Established guidelines and recommendations for high-quality research formulating joint standards could promote the adoption of the display technologies and facilitate exchange among researchers, educators and developers with widely different experiences and approaches.[57]

Similar to the words of David A. Cook, professor of medicine and medical education, we suggest placing more emphasis on the 'How' and 'When' to use AR and MR-based learning and to focus less on 'Whether'.[55] Answering these questions researchers, educators and developers should share and evaluate the instructional design and learning theory-based methods while looking into effective use of simulation, and integration of the display technologies within and between institutions. Eventually, this could also provide an understanding of learning concepts revealed from the included studies involving intrinsic benefits of motivation, physical interaction activating kinesthetic schemes, skill retention, transferability of simulation confidence, mobile learning and using oneself as a learning object. By defining instructional objectives beforehand, the display technologies should be used only when it could refine or even replace training programs and curricula.

With that being said partially immersive environments such as AR and MR may offer unique qualities for specifically, assessment and training procedural strategies integrating real patient data and without breaching patient safety. By using non-invasive sensors for imaging, the display technologies could complement the established imaging technologies of MRI, CT scan and ultrasound for monitoring of technical performance with an objective-comparative function as observed in our review.[27,29,50] To tap the full potential of the display technologies, the study and application design must be based on a throughout investigation of the educational context, learner types and learning objectives whether the latter being cognitive, technical, or non-technical such as measuring situational awareness, communication, or stress coping.

Conclusions

This review reports the current state of AR and MR-based applications for healthcare education beyond surgery. Studies based on both display technologies across various specialties and subjects states an increased number of established applications moving the research base away from feasibility studies on prototypes. All included studies suggested various healthcare educational benefits by the display technologies which significantly outperformed traditional learning approaches in 11 studies, specifically regarding the acquisition of anatomy knowledge and needle insertion skills. Yet, this review identifies multiple shortcomings of the studies. Study

quality was low-to-medium especially due to lacking validity of the evaluation instruments, heterogeneity of research designs and widely varied reporting. Future studies are thus needed for researchers, educators and developers to build an evidence base defining suitable research designs and instructional objectives achievable by AR and MR-based applications, for these to complement conventional learning, curricula, and conduct a transformation in healthcare education.

Acknowledgements

We would like to thank for financial support by the institutional funds of the Copenhagen Academy of Medical Education and Simulation (CAMES), and valuable feedback by the employees of the academy.

Conflict of Interest

PD holds a professorship with the University of Stavanger, Norway that is supported unconditionally by a grant from the Laerdal Foundation in Norway.

References

1. Kamphuis C, Barsom E, Schijven M, Christoph N. Augmented reality in medical education? Perspect Med Educ. 2014;3(4):300-11.
2. Chen L, Day TW, Tang W, John NW. Recent developments and future challenges in medical mixed reality. Proceedings of 16th IEEE International Symposium on Mixed and Augmented Reality (ISMAR); 9-13 October 2017. Nantes, France: IEEE; 2017.
3. Silva JNA, Southworth M, Raptis C, Silva J. Emerging applications of virtual reality in cardiovascular medicine. JACC Basic Transl Sci. 2018;3(3):420-30.
4. Barsom EZ, Graafland M, Schijven MP. Systematic review on the effectiveness of augmented reality applications in medical training. Surg Endosc. 2016;30(10):4174-83.
5. Azuma R, Baillot Y, Behringer R, Feiner S, Julier S, MacIntyre B. Recent advances in augmented reality. IEEE Comput Grap Appl. 2001; 21(6): 34-47.
6. Milgram P, Takemura H, Utsumi A, Kishino F. Augmented reality: a class of displays on the reality-virtuality continuum. Telemanipulator and Telepresence Technologies. 1995;2351:282-92.
7. Milgram P, Kishino F. Taxonomy of mixed reality visual displays. IEICE Transactions on Information Systems. 1994; E77-D(12): 1321-9.
8. Brigham TJ. Reality check: basics of augmented, virtual, and mixed reality. Med Ref Serv Q. 2017;36(2):171–8.
9. Zhu E, Lilienthal A, Shluzas LA, Masiello I, Zary N. Design of mobile augmented reality in health care education: a theory-driven framework. JMIR Med Educ. 2015;1(2):e10.
10. Wojtczak A. Glossary of medical education terms: part 4. Med Teach. 2002;24(5):567–8.
11. Smith ML, Foley MR. Transforming clinical education in obstetrics and gynecology: gone is the day of the sage on the stage. Obstet Gynecol. 2016;127(4):763–7.
12. Ellis H, Abdalla S. A history of surgery. Boca Raton: CRC Press; 2018.
13. Zhu E, Hadadgar A, Masiello I, Zary N. Augmented reality in healthcare education: an integrative review. PeerJ. 2014;2:e469.
14. Yoon JW, Chen RE, Kim EJ, Akinduro OO, Kerezoudis P, Han PK, et al. Augmented reality for the surgeon: systematic review. Int J Med Robot. 2018;14(4):e1914.
15. Aggarwal A. The evolving relationship between surgery and medicine. Virtual Mentor. 2010;12(2):119–23.
16. Martin Elizabeth A EA. Concise medical dictionary. Oxford: Oxford University Press; 2015.
17. Wong K, Yee HM, Xavier BA, Grillone GA. Applications of augmented reality in otolaryngology: a systematic review. Otolaryngol Head Neck Surg. 2018;159(6):956–67.
18. Whittemore R. Combining evidence in nursing research: methods and implications. Nurs Res. 2005;54(1):56–62.
19. Whittemore R, Knafl K. The integrative review: updated methodology. J Adv Nurs. 2005;52(5):546–53.
20. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J. 2009;26(2):91–108.
21. TED. TED: Ideas worth spreading. [Cited 1 September 2018]: Available from: https://www.ted.com.
22. iTunes Podcast app. Podcasts Downloads on iTunes. [Cited 1 September 2018]; Available from: https://itunes.apple.com/gb/genre/podcasts/id26.
23. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. JAMA. 2007;298(9):1002–9.
24. Cook DA, Reed DA. Appraising the quality of medical education research methods: the medical education research study quality instrument and the newcastle-ottawa scale-education. Acad Med. 2015;90(8):1067–76.
25. Côté L, Turgeon J. Appraising qualitative research articles in medicine and medical education. Med Teach. 2005;27(1):71–5.
26. Stacy R, Spencer J. Assessing the evidence in qualitative medical education research. Med Educ. 2000;34(7):498–500.
27. Abhari K, Baxter JSH, Chen ECS, Khan AR, Peters TM, de Ribaupierre S, et al. Training for planning tumour resection: augmented reality and human factors. IEEE Trans Biomed Eng. 2015;62(6):1466–77.
28. Bifulco P, Narducci F, Vertucci R, Ambruosi P, Cesarelli M, Romano M. Telemedicine supported by augmented reality: an interactive guide for untrained people in performing an ECG test. Biomed Eng Online. 2014;13:153.
29. Jeon Y, Choi S, Kim H. Evaluation of a simplified augmented reality device for ultrasound-guided vascular access in a vascular phantom. J Clin Anesth. 2014;26(6):485–9.
30. Kugelmann D, Stratmann L, Nühlen N, Bork F, Hoffmann S, Samarbarksh G, et al. An augmented reality magic mirror as additive teaching device for gross anatomy. Ann Anat. 2018;215:71–7.
31. Ma M, Fallavollita P, Seelbach I, Von Der Heide AM, Euler E, Waschke J, et al. Personalized augmented reality for anatomy education. Clin Anat. 2016;29(4):446–53.
32. Mewes A, Heinrich F, Kägebein U, Hensen B, Wacker F, Hansen C. Projector-based augmented reality system for interventional visualization inside MRI scanners. Int J Med Robot. 2019;15(1):e1950.
33. Solbiati M, Passera KM, Rotilio A, Oliva F, Marre I, Goldberg SN, et al. Augmented reality for interventional oncology: proof-of-concept study of a novel high-end guidance system platform. Eur Radiol Exp. 2018;2:18.
34. Sutherland C, Hashtrudi-Zaad K, Sellens R, Abolmaesumi P, Mousavi P. An augmented reality haptic training simulator for spinal needle procedures. IEEE Trans Biomed Eng. 2013;60(11):3009–18.
35. Wang LL, Wu HH, Bilici N, Tenney-Soeiro R. Gunner goggles: implementing augmented reality into medical education. Stud Health Technol Inform. 2016;220:446–9. PMID: 27046620
36. Ferrer-Torregrosa J, Torralba J, Jimenez MA, García S, Barcia JM. ARBOOK: Development and assessment of a tool based on augmented reality for anatomy. J Sci Educ Technol. 2015;24:119–24.
37. Ferrer-Torregrosa J, Jiménez-Rodríguez MÁ, Torralba-Estelles J, Garzón-Farinós F, Pérez-Bermejo M, Fernández-Ehrling N. Distance learning ects and flipped classroom in the anatomy learning: comparative study of the use of augmented reality, video and notes. BMC Med Educ. 2016;16(1):230.
38. Huang CY, Thomas JB, Alismail A, Cohen A, Almutairi W, Daher NS, et al. The use of augmented reality glasses in central line simulation: "see one, simulate many, do one competently, and teach everyone". Adv Med Educ Pract. 2018;9:357–63.
39. Küçük S, Kapakin S, Göktaş Y. Learning anatomy via mobile augmented reality: Effects on achievement and cognitive load. Anat Sci Educ. 2016;9(5):411–21.
40. Leitritz MA, Ziemssen F, Suesskind D, Partsch M, Voykov B, Bartz-Schmidt KU, et al. Critical evaluation of the usability of augmented reality ophthalmoscopy for the training of inexperienced examiners. Retina. 2014;34(4):785–91.
41. Moro C, Štromberga Z, Raikos A, Stirling A. The effectiveness of virtual and augmented reality in health sciences and medical anatomy. Anat Sci Educ. 2017;10(6):549–59.

42. Noll C, von Jan U, Raap U, Albrecht U-V. Mobile augmented reality as a feature for self-oriented, blended learning in medicine: randomized controlled trial. JMIR Mhealth Uhealth. 2017;5(9):e139.

43. Rai AS, Rai AS, Mavrikakis E, Lam WC. Teaching binocular indirect ophthalmoscopy to novice residents using an augmented reality simulator. Can J Ophthalmol. 2017;52(5):430–4.

44. Rochlen LR, Levine R, Tait AR. First-person point-of-view-augmented reality for central line insertion training: a usability and feasibility study. Simul Healthc. 2017;12(1):57–62.

45. Siebert JN, Ehrler F, Gervaix A, Haddad K, Lacroix L, Schrurs P, et al. Adherence to AHA guidelines when adapted for augmented reality glasses for assisted pediatric cardiopulmonary resuscitation: a randomized controlled trial. J Med Internet Res. 2017;19(5):e183.

46. Wang S, Parsons M, Stone-McLean J, Rogers P, Boyd S, Hoover K, et al. Augmented reality as a telemedicine platform for remote procedural training. Sensors. 2017;17(10):2294.

47. Zhu E, Fors U, Smedberg Å. Exploring the needs and possibilities of physicians' continuing professional development - an explorative qualitative study in a chinese primary care context. PLoS One. 2018;13(8):e0202635.

48. Aebersold M, Voepel-Lewis T, Cherara L, Weber M, Khouri C, Levine R, et al. Interactive anatomy-augmented virtual simulation training. Clin Simul Nurs. 2018;15:34–41.

49. Albrecht U-V, Folta-Schoofs K, Behrends M, von Jan U. Effects of mobile augmented reality learning compared to textbook learning on medical students: randomized controlled pilot study. J Med Internet Res. 2013;15(8):e182.

50. Keri Z, Sydor D, Ungi T, Holden MS, McGraw R, Mousavi P, et al. Computerized training system for ultrasound-guided lumbar puncture on abnormal spine models: a randomized controlled trial. Can J Anaesth. 2015;62(7):777–84.

51. Moult E, Ungi T, Welch M, Lu J, McGraw RC, Fichtinger G. Ultrasound-guided facet joint injection training using Perk Tutor. Int J Comput Assist Radiol Surg. 2013;8(5):831–6.

52. Robinson AR, Gravenstein N, Cooper LA, Lizdas D, Luria I, Lampotang S. A mixed-reality part-task trainer for subclavian venous access. Simul Healthc. 2014;9(1):56–64.

53. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.

54. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach. 2005;27(1):10–28.

55. Cook DA. Where are we with web-based learning in medical education? Med Teach. 2006;28(7):594–8.

56. Jensen L, Konradsen F. A review of the use of virtual reality head-mounted displays in education and training. Education and Information Technologies. 2018;23(4):1515–29.

57. Cheng A, Kessler D, Mackinnon R, Chang TP, Nadkarni VM, Hunt EA, et al. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. Simul Healthc. 2016;11(4):238–48.

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 posttests 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)	Mixed methods study: Three-group RCT and survey	Test: 1) Self-developed multiple choice Questionnaire:	No significant difference in test scores between the three learning modes ($p<.874$). Adverse effects as dizziness were significantly	Self-developed for mobile devices (10 minutes)	AR	13.5	5

	(Learning of skull anatomy)	(59 health science students, tablet/Control=22; VR=20; AR=17)	2) Scale on adverse health effects 3) Likert scale on LE	experienced in the VR group vs. AR and tablet group (p<.001). Perception of AR was high but not significant				
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 (Simulating subclavian venous access (SCVA/CVC) without US-guidance on phantom)	Mixed methods study: Three-group non-randomized comparison and survey (65 physicians of different training categories, novices=25; intermediates=24; experts=16)	Test (pre- and post-intervention without assistive technology): 1) SCVA score 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)	All participants significantly improved SCVA score ($p<.0001$) and time ($p<.0001$). The participants 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 ($p<.0001$)	Self-developed for computer with tracker sensor-recognizing a virtual model of the phantom registered within a 3D-printed phantom built-up of head and thorax CT scan (Until users felt confident)	MR	13.5	7
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

Appendix 2.

Distribution of studies across medical specialty or health science, number of studies, and participants enrolled according to number of studies

Medical Specialty or Health Science	Subjects of Healthcare Education	No. Studies	No. Participants (according to number of studies)
Anatomy (6 studies)	Foot Muscles	1	171
	Lower Limb	1	211
	Skull	1	59
	Human Gross Anatomy	2	880+72
	Neuroanatomical Pathways	1	70
Anesthesia (7 studies)	Central Vein Catheterization	4	32+20+65+40
	Lumbar Puncture	1	24
	Spinal Needle Insertion	1	10
	Ultrasound Examination for Trauma (Telemedicine)	1	24
Cardiology	Electrocardiogram Recording	1	20
Dermatology	Skin Diseases	1	44
Family Medicine	Continuing Professional Development*	1	15
Forensic Medicine	Gunshot Wounds	1	10
Gastroenterology	Nasogastric Tube Insertion (Nursing)	1	69
Neurology	Shelf Exam Preparation	1	24
Ophthalmology (2 studies)	Binocular Indirect Ophtalmoscopy	2	37+31
Orthopedics	Facet Joint Injections	1	26
Pediatrics	Cardiopulmonary Resuscitation	1	20
Radiology (3 studies)	Resection Planning (Neurosurgery)	1	21
	Needle Insertion (MRI)	1	8
	Needle Insertion (CT scan)	1	Not Specified

*A theoretical application was devised from an anatomy aplication.[47]