

Beyond technological immersion: augmented and mixed reality in undergraduate medical education. A critical review of comparative studies.

Más allá de la inmersión tecnológica: realidad aumentada y realidad mixta en la educación médica de grado. Revisión crítica de estudios comparativos.

Lauricia Jacqueline Macías Mendoza¹, Yoiler Batista Garcet¹, Lissette Beatriz Villavicencio Cedeño²

¹ San Gregorio University of Portoviejo, School of Medicine, Portoviejo, Manabí, Ecuador. ljmacias@sangregorio.edu.ec, <https://orcid.org/0000-0002-8779-9091>; ybatista@sangregorio.edu.ec, <https://orcid.org/0000-0002-7851-5763>. ² National University of Tumbes, Graduate School, PhD in Applied Mathematical Statistics, Tumbes, Peru. iso_villa@hotmail.com, <https://orcid.org/0000-0001-8475-3947>

* Corresponding author: ljmacias@sangregorio.edu.ec

Received: 30/9/26; Accepted: 27/10/26; Published: 29/10/26

Summary

Introduction: Augmented reality (AR) and mixed reality (MR) allow the integration of digital information with the physical environment, but their educational utility depends on the learning outcomes they achieve. **Objective:** To critically analyze comparative studies on AR/MR in undergraduate medical studies, considering the competency assessed and the instructional design. **Methods:** A critical review with narrative synthesis of a corpus identified through previous reviews and progressive searches in PubMed between January and September 2026, with a cutoff date of September 29, 2026, was conducted. Knowledge, spatial understanding, procedural performance, retention, and transfer were examined; student experience was treated as a secondary outcome. Directly relevant reports and supplementary evidence were differentiated by population or access to documentation. **Results:** Eighteen reports published between 2017 and 2026 were examined (16 directly relevant and two contextual). Benefits in certain anatomical and procedural tasks coexisted with comparisons showing no significant differences. Immediate improvements were not maintained in all follow-up studies. In point-of-care studies, the non-inferiority of specific measurements was sensitive to analytical exclusions; lumbar puncture did not demonstrate greater clinical success. **Conclusions:** The evidence reviewed suggests specific benefits, without establishing overall superiority of augmentation/reduction. The correlation between resource, activity, and competence constitutes a plausible pedagogical explanation. The non-exhaustive coverage, study limitations, and the absence of a comprehensive formal assessment of bias and certainty restrict curricular recommendations.

Keywords: augmented reality; mixed reality; medical education; medical students; anatomy; clinical competence; critical review.

Resumen

Introducción: La realidad aumentada (RA) y la realidad mixta (RM) permiten integrar información digital con el entorno físico, pero su utilidad educativa depende de los resultados de aprendizaje que alcanzan. **Objetivo:** Analizar críticamente estudios comparativos sobre RA/RM en Medicina de grado, atendiendo a la competencia evaluada y al diseño instruccional. **Métodos:** Revisión crítica

con síntesis narrativa de un corpus identificado mediante revisiones previas y búsquedas progresivas en PubMed entre enero y septiembre de 2026, con fecha de corte el 29 de septiembre de 2026. Se examinaron conocimiento, comprensión espacial, desempeño procedimental, retención y transferencia; la experiencia del estudiante se trató como resultado secundario. Se diferenciaron informes directamente pertinentes y evidencia complementaria por población o acceso documental.

Resultados: Se examinaron 18 informes publicados entre 2017 y 2026 (16 directamente pertinentes y dos de carácter contextual). Los beneficios en determinadas tareas anatómicas y procedimentales coexistieron con comparaciones sin diferencias significativas. Las mejoras inmediatas no se mantuvieron en todos los seguimientos. En POCUS, la no inferioridad de mediciones específicas fue sensible a exclusiones analíticas; en punción lumbar no se demostró mayor éxito clínico.

Conclusiones: La evidencia examinada sugiere beneficios específicos, sin establecer superioridad general de la RA/RM. La correspondencia entre recurso, actividad y competencia constituye una explicación pedagógica plausible. La cobertura no exhaustiva, las limitaciones de los estudios y la ausencia de una evaluación formal completa de sesgo y certeza restringen las recomendaciones curriculares.

Palabras clave: realidad aumentada; realidad mixta; educación médica; estudiantes de Medicina; anatomía; competencia clínica; revisión crítica.

1. Introduction

Augmented reality (AR) overlays digital information onto the real-world environment, while mixed reality (MR) allows for a more interactive integration between virtual objects and physical space. In medical education, both technologies have been used to represent three-dimensional anatomy, guide procedures, train surgical skills, and support decision-making. Their pedagogical appeal stems from a possibility difficult to replicate with two-dimensional materials: simultaneously maintaining the real-world context and adding layers of visual, spatial, or procedural information. Extended reality (XR) is used as an umbrella term; its designation alone is insufficient to deem an intervention eligible, as it must effectively incorporate AR or MR (1, 2). Interest in these technologies has grown rapidly, but the evidence is not uniform. A 2022 meta-analysis, based on 13 studies and 654 participants, found improvements in performance time, confidence, and satisfaction, but did not demonstrate consistent benefits in knowledge or skills. A more recent synthesis, published in 2025 and comprising 29 publications, reported higher scores in skills, a lower failure rate, and shorter execution times, without a statistically significant advantage in knowledge. However, heterogeneity and potential publication bias necessitate interpreting these effects with caution (3, 4).

In anatomy, Williams et al. identified 17 comparative studies of AR/MRI; 12 of these showed predominantly no significant differences compared to other methods. The syntheses by García-Robles et al. and Moro et al. cover broader technologies and populations, so their findings should not be automatically extrapolated to AR/MRI in undergraduate medical education. Mahmud et al. examined applications for clinical skills using a scope review design (5-8). In the Spanish Journal of Medical Education, González Salgado et al. addressed design criteria for virtual environments for anatomy, and Ampuero Valenzuela et al. analyzed tools for teaching morphology. These works provide pedagogical context but do not replace specific comparative evidence on AR/MRI (9, 10).

This review critically analyzes the results of AR/RM in undergraduate medical studies and their application conditions. The central question is for which competency, against which comparator, and with which instructional design benefits are observed. The contribution lies in distinguishing between immediate acquisition, retention, and transfer, and in examining the contribution of the accompanying instructional materials. The relationships identified constitute interpretive explanations, not causal evidence of pedagogical moderators.

2. Methods

2.1. Design and scope

A critical review of comparative studies on AR/MR in undergraduate medical education was conducted, with a narrative synthesis organized by competencies and assessment points. The literature search was carried out progressively between January and September 2026, and the reports examined were published between 2017 and 2026. Results, comparators, and limitations were compared, without estimating an overall effect or claiming exhaustiveness. Since complete search histories, filters, and independent selection were not maintained, an up-to-date systematic review is not considered; therefore, the work is classified as a critical review of comparative studies. The presentation of the synthesis was guided by SWiM, without declaring its full application (11). Supplementary materials S1, S1A, and S2–S5 constitute a retrospective organization of the preserved information, not a contemporaneous record of a systematic procedure.

2.2. Review Question

The question was formulated using PICOS: In undergraduate medical students (P), do AR- or RM-based educational interventions (I), compared with conventional teaching, 2D/3D resources, standard simulation, or other active methodology (C), improve knowledge, spatial understanding, procedural performance, retention, transfer, cognitive load, or related educational outcomes (O), in comparative experimental or quasi-experimental studies (S)?

2.3. Sources and search strategy

Three periods were identified: the progressive execution of the literature search (January–September 2026), the publication period of the reports examined (2017–2026), and the phase of updating recent publications (2024–2026). The bibliographic closing date was September 29, 2026; this should not be confused with the individual execution date of each search.

Historical identification relied on the lists of primary studies from the reviews by Baashar et al. (3), Zhang et al. (4), and Williams et al. (5). These secondary works were used for locating and contextualizing the literature, not as additional units of analysis to the primary studies. Bibliographic updates were conducted using PubMed/MEDLINE, accessed via PubMed. PubMed Central and indexed publisher websites were primarily used to retrieve full texts and verify information. References from relevant publications were also examined. The historical identification and updating process was subjected to the relevance criteria described in section 2.4.

A PubMed equation consisting of three blocks joined by AND (technology, population, and educational setting) is preserved, with alternative terms linked by OR, searches in Title/Abstract fields, and truncation in hologram*. Supplementary material S1 reproduces the preserved PubMed equation and distinguishes between declared relevance rules and automated search filters, whose historical application cannot be verified. Run times per source, complete filter histories, export files, and original counts of results and duplicates are not preserved. Therefore, the equation can be reused, but it is not possible to retrospectively reconstruct a comprehensive search or validate the counts in the flowchart presented in a previous version using primary records. These limitations justify the adopted designation of critical review; historical counts are documented, without evidentiary value, only in S1A.

2.4. Relevance criteria and outcomes

- Target population: undergraduate medical students. Reports with mixed populations and no separable outcomes are presented solely for educational purposes; they do not support specific conclusions regarding efficacy in medicine.

- Intervention: AR or MR with contextual overlay or integration of digital content with the physical environment. Direct student use was distinguished from instruction delivered via a viewer used by the instructor. The term XR, by itself, does not establish the presence of an AR/MR component.
- Comparator: conventional teaching, atlas, two-dimensional or three-dimensional resources, standard simulation, teacher instruction, or other active educational strategy.
- Primary outcomes of the synthesis: objectively measured knowledge, spatial understanding, and procedural performance; retention in a delayed assessment; and transfer to a task or context other than training. Secondary outcomes: satisfaction, motivation, confidence, usability, and cognitive load. This hierarchy organizes the analysis and is not presented as pre-specified in a protocol. OSCE and OSPE correspond to objective structured clinical and practical assessments, respectively.
- Design: randomized trial, prospective comparative study or quasi-experimental study with comparator group.
- Delimitation: Virtual reality without an AR/MR component, training exclusively for residents or professionals, and studies without comparisons did not provide direct evidence to answer the question. Previous reviews and reports with population or access limitations were used for context, distinct from directly relevant comparisons.

2.5. Selection, extraction and critical evaluation

The search and selection of publications are attributed to all three authors, according to the statement of contributions. Data extraction and curation are attributed to MM and VC (two authors), and analysis and interpretation to all three authors. These figures describe their overall involvement, not the number of independent reviewers at each stage. The retained documentation does not allow for verification of duplicate screening, independent extraction, separate bias assessment, consensus procedures, or peer review by a third reviewer. These procedures are not retrospectively attributed to the study.

Titles and abstracts of potentially relevant studies were examined, and the primary publication was checked, where available, to identify population, intervention, comparator, design, outcomes, and results. The original phrase “or through sufficient bibliographic information” was removed due to its ambiguity: when the full report was unavailable (e.g., Xiao et al.), the information was limited to what was reported in the original abstract and presented as context, without being considered comprehensive verification of eligibility or sufficient support for an inference of efficacy. Mixed samples without separable analyses for medical students (17) are also distinguished from the directly applicable body of work. Individual inclusion/exclusion decisions and the resolution of discrepancies lack a retrievable historical record; this lack of traceability is declared as a limitation of the work. Supplement S1 and its section S1A detail the traceability limitations; S2 presents individual critical fact sheets, and S3 describes the synthesis criteria. None of these documents document unrecorded procedures.

The critical narrative appraisal focused on participant allocation and confounding, comparator adequacy, analytical biases and exclusions, missing data, instrument properties, and the selection or timing of results. The individualized matrix in Supplement S4 presents unverifiable information and aspects of each report, taking into account the specific characteristics of the described design. These aspects are conceptually related to RoB 2 and ROBINS-I (12, 13), but do not constitute formal judgments on the results obtained with their signaling questions and algorithms. For this reason, low, moderate, or high risk of bias categories are not assigned, nor is an independent procedure that cannot be verified presented as completed.

2.6. Synthesis Strategy

A separate meta-analysis was not conducted due to differences in interventions, comparators, instruments, and assessment timing. The synthesis was organized by competency and assessment time (immediate acquisition, retention, and transfer), and favorable, non-significant, and non-inferiority outcomes were explicitly contrasted. Effects reported in previous meta-analyses are provided for context, without being incorporated as new observations or including new primary studies. The number of studies with $p < 0.05$ was not used as an estimate of effectiveness. The narrative reading gave greater support to objective results from relevant populations and full reports, and less interpretive weight to summaries, non-separable samples, pilot designs, or findings dependent on exclusions. This classification is not a quantitative weighting or a formal assessment of certainty. The PRISMA flow diagram was removed from the main version because its historical counts are not auditable; Appendix S1A retains only its retrospective review.

2.7. Confidence in the findings

Interpretive confidence was examined by outcome, considering population relevance, access to the primary report, comparator, objective measurement, consistency of results, precision, follow-up duration, and design limitations. Full reports with objective outcomes and relevant comparators provided direct support for the specific tasks studied; pilot studies, those with analytical exclusions, or those with exclusively immediate assessments were interpreted with greater reservations. Reports based solely on summaries or on inseparable mixed samples served a contextual function. Table 2 and Supplement S5 illustrate this qualitative hierarchy, which is not equivalent to GRADE and does not allow for the assignment of formal levels of certainty per outcome or the formulation of graded recommendations.

3. Results

3.1. Reports examined and general characteristics

The corpus examined comprises 18 reports published between 2017 and 2026: 16 comparisons directly relevant to undergraduate medical studies and two contextual reports (17, 20). Ten of these were published up to 2023, and eight were published between 2024 and 2026. This temporal division does not constitute evidence of the individual location pathway in historical or updated searches, which lack a verifiable, complete export. The research focused on anatomy, neuroanatomy, histology, and procedural training. Table 1 describes the documentary corpus and its limitations, not an exhaustive selection reconstructed using a PRISMA workflow. Zeedzen-Scheffers et al. (17) is considered contextual because it includes Medicine and Biomedical Sciences without separable results, and Xiao et al. (20) is identified as evidence available only in the primary abstract. Supplement S1-S1A documents the traceability strategy and limitations. S2 contains critical reports, S3 explains the synthesis criteria, S4 documents methodological observations, and S5 presents the qualitative synthesis by outcome.

Table 1. Characteristics and results of the 18 reports examined (16 relevant and 2 contextual).

Study (Ref)	n	Area	Comparison	Main finding
Noll (29)	44	Dermatology	Mobile AR vs. conventional teaching	Mobile AR promoted autonomous learning; small sample size, no evidence of clinical transfer.
Stojanovska (15)	64	Anatomy	HoloLens vs. cadaveric dissection	No significant differences compared to dissection; no equivalence is demonstrated.
Schoeb (21)	164	Bladder catheterization	RM/HoloLens vs. instructor	Higher OSCE score with RM (21.49 vs 19.96); task-specific assessment.
Weeks (14)	30	Head and neck anatomy	Holographic AR vs. 2D screen	Best immediate post-test with RA (effect size: 0.73); retention not evaluated.
Pickering (16)	200*	Neuroanatomy	RM vs. active screencast	Similar gains; some measures favored screencasting. Variable denominators.
Farshad-Amaker (25)	44	Ultrasound	RA-US vs. US standard	There was no uniform RA advantage; possible influence of order on cross-design.
Guha (22)	36	Surgical skills	HoloLens vs. traditional video/teaching	Signs of technical improvement with MRI; pilot study with blinded evaluators.
Hayasaka (24)	30	Epidural anesthesia	AR/MRI vs. conventional technique	Greater accuracy in epidural simulator; three groups of ten participants.
Malik (30)	56	Otolaryngology	Transmission with HoloLens 2 from the teacher vs. conventional remote case	Greater gain in knowledge (12 versus 9 points); subsequent exclusions.
D'Aiello (31)	59	Cardiac Anatomy	RM vs. holographic videos vs. slides	Anatomical content-dependent effects; exploratory comparison.
Zeedzen-Scheffers (17)	28*	Neuroanatomy	RA vs. anatomical atlas	No significant difference in knowledge; non-separable mixed sample (contextual).
Sarkar (26)	34	Lumbar puncture	AR + simulation vs. standard simulation	No significant technical differences; greater self-reported confidence.
Hadida Barzilai (23)	21	Mastoidectomy	Guided AR vs. traditional training	Highest milling score (19.5/25 vs. 12/25); simulated task.
Zingg (18)	48	Anatomy	Adaptive holographic AR vs. traditional resources	Better theoretical and practical results; multi-component intervention.
Felten (27)	188*	Clinical lumbar puncture	AR + haptics vs. bedside teaching	No superiority in clinical success; 53 procedures analyzed per protocol.
Mulder (28)	56	Cardiac POCUS	XR/RA vs. instructor	Non-inferiority in specific measurements; EPSS sensitive to exclusions.

Xiao (20)	152	Histology	AR vs. traditional microscopy	Better retention and OSPE according to the summary; contextual information not fully verified.
Starszak (19)	236	Anatomy	RM vs. 3D printing vs. control	Immediate improvement with RM, with no significant difference in delayed retention.

Notes. Pickering: The text states 200 students with pretest and posttest, but the results table presents 127 + 75 for multiple-choice questions and 84 + 62 = 146 for short answers and overall score; this discrepancy from the primary report is retained without inferring a single denominator. Hayasaka: 30 participants, three groups of 10. d'Aiello: 59 participants, groups of 20, 20, and 19. Felten: 188 participating students and 53 procedures analyzed per protocol, units that are not interchangeable. An aggregate total sample is not calculated. Zeedzen-Scheffers (17) and Xiao (20) are contextual only and are not used to support specific inferences of effectiveness in undergraduate Medicine.

3.2. Anatomy, histology and spatial understanding

Anatomy results were not uniform. Weeks et al. reported better immediate results with holographic representation than with a 2D screen. Stojanovska et al. observed no significant differences in the compared practical assessments; their design combined cohorts and assessment modalities. Pickering et al. used a quasi-randomized design and found advantages of screencasting in multiple-choice questions, with no significant differences in other comparisons (14-16). For context, Zeedzen-Scheffers et al. studied 28 participants from Medicine and Biomedical Sciences: they observed no significant differences in knowledge gain. The sample size does not allow for separate inference for Medicine (17). In TEACHANATOMY, 48 students were assigned to adaptive holographic AR with gamification or traditional learning. The medians were 18.8 vs. 9.4 on the theory test and 14.0 vs. 5.0 on the practical test ($p < 0.001$). The contrast evaluates a set of educational components, not the isolated effect of the device (18). Starszak et al. included 236 students: MRI obtained an immediate advantage over the control, but no significant differences were observed in delayed retention (19). The abstract by Xiao et al. describes favorable results of RA in histology in 152 students, with follow-up at 10 weeks; as the full text is not available, it is considered supplementary information and random assignment is not assumed (20).

3.3. Procedural and surgical skills

Schoeb et al. reported a higher OSCE score for bladder catheterization with MRI (21). Guha et al., in 36 students with blinded evaluators, observed a greater gain in dexterity for arteriotomy and closure compared to a video tutorial (22). In mastoidectomy, Hadida Barzilai et al. found reaming scores of 19.5/25 versus 12/25 ($p < 0.01$), assessed by two blinded otologists (23). Hayasaka et al. studied 30 participants and observed improvements in accuracy using an epidural anesthesia simulator (24). These tasks assess performance under specific conditions and do not, by themselves, demonstrate transfer to patients. In the crossover study by Farshad-Amacker et al., puncture time with augmented ultrasound did not differ significantly from conventional puncture time (25). In the pilot study by Sarkar et al. No significant differences were identified in the main technical outcomes of lumbar puncture, despite a favorable subjective experience (26). Felten et al. examined clinical transfer: of 188 participating students, 53 procedures were analyzed per protocol. Success was 13/28 (46.4%) versus 10/25 (40.0%); the results section reports $p = 0.85$. The trial was stopped early due to futility. These findings do not establish equivalence or rule out clinically relevant differences with sufficient precision (27). Mulder et al. included 56 students and compared HoloLens 2 training versus conventional instruction for point-of-care cardiac ultrasound (POCUS). They reported non-inferiority for E-point septal separation (EPSS) and tricuspid annular plane systolic excursion (TAPSE), with margins of 3 and 4 mm, respectively. However, they excluded two measurements per outcome; For EPSS, the inclusion of extreme values from the XR group prevented statistical

confirmation of non-inferiority. Interpretation should be limited to these outcomes and analyses. The design does not constitute a cost-effectiveness evaluation nor does it demonstrate institutional scalability (28).

3.4. Knowledge, retention and transfer

The results for declarative knowledge were heterogeneous. As an external context, Zhang et al. reported a weighted mean difference of 2.92 points (95% CI -1.73 to 7.57), with no statistical evidence of superiority. The range includes no effect and possible benefits; it does not demonstrate equivalence. Their aggregate skills score showed high heterogeneity ($I^2 = 95.8\%$), which limits the interpretation of a single magnitude across different tasks (4). Longitudinal assessment was less abundant than immediate assessment. Starszak et al. found no significant differences at follow-up (19). The favorable result at 10 weeks reported in the summary by Xiao et al. provides additional evidence, but is insufficient to establish the overall persistence of the benefit (20). Felten et al. assessed procedures in patients, but the number of procedures and the discontinuation due to futility limit the precision (27). Inferences about retention and transfer must distinguish between outcome, follow-up, and the unit analyzed; they cannot be extrapolated from an immediate post-test.

3.5. Cognitive load, motivation and student experience

The student experience was frequently favorable, although not uniformly so. Noll et al. evaluated mobile AR in blended self-directed learning (29). Malik et al. analyzed 56 students: the mean gain on a 20-point questionnaire was 12 versus 9 points. The viewers were used by instructors, and students received the feed via Microsoft Teams; more questions were asked, but overall measured engagement did not increase significantly. Post-assignment exclusions and group imbalance limit interpretation (30). d'Aiello et al. evaluated 59 students in three modalities and found content-dependent effects (31). Acceptance does not replace the demonstration of clinical competence.

3.6. Methodological limitations of the studies

The limitations were analyzed study by study and linked to the interpreted outcome. Pickering's quasi-random assignment (16), denominator variation, and post-assignment exclusions in Malik (30) limit internal comparability; Farshad-Amacker's crossover design (25) requires attention to order and prior learning. The blinding of evaluators reported by Guha (22) and Hadida Barzilai (23) strengthens the measurement of certain simulated tasks, although not their clinical generalizability. In Mulder (28), inferences about EPSS are sensitive to the exclusion of values; in Felten (27), futility termination and the 53 procedures analyzed restrict the precision regarding transfer. Studies based solely on abstracts do not allow for the verification of all relevant domains. The S4 matrix differentiates observable limitations from unverified information, without converting them into formal RoB 2/ROBINS-I classifications. The small size primarily affects accuracy, and the absence of student blinding does not automatically demonstrate a high risk in all outcomes.

Table 2. Interpretive scope of the evidence by outcome.

Outcome	Observed pattern	Main constraint	Admissible inference
Spatial knowledge and understanding	Improvements in specific tasks and comparisons with no differences.	Comparators and heterogeneous instruments; incomplete allocation and data in some studies.	No overall superiority is demonstrated.
Procedural performance	Benefits in catheterization, arteriotomy, epidural or	Small samples and simulated tasks; different	Do not extrapolate to

	mastoidectomy; variable results in other tasks.	assistance during the test.	global clinical competence.
Retention	Anatomical follow-up showed no differences; favorable signal in histology reported only in summary.	Few follow-ups, different timeframes, and limited documentary access in histology.	Persistence of the effect uncertain.
Transfer and POCUS	Lumbar puncture without superiority; POCUS with non-inferiority sensitive to exclusions.	Different analytical units; futility and sensitivity to excluded values.	Conclusions limited to each result.
Experience and costs	Frequently favorable acceptance; without its own economic synthesis.	Self-reporting and potential novelty; costs not assessed in a comparable manner.	Do not infer effectiveness, savings, or sustainability.

4. Discussion

The synthesis shows favorable results in certain anatomy and procedural tasks, along with comparisons showing no significant advantage. The most cautious interpretation is that the RA/RM value depends on the educational problem and the application conditions. The alignment between visualization, activity, and competence is a plausible explanation; however, it was not subjected here to causal testing or a meta-regression analysis. The differences between favorable results and those showing no statistically significant advantage may depend on the comparator (conventional instruction versus active resources), the intensity and duration of the activity, the student's prior experience, the feedback structure, and the proximity between the practiced task and the assessment instrument. TEACHANATOMY (18) integrates adaptation and gamification into the intervention; Guha (22) incorporates guided procedural practice and blinded assessment; Pickering (16) uses an active comparator integrated into the curriculum. A separate moderator analysis is not available to isolate the effect of the device from these components. Improvement during a guided task also does not equate to competence maintained unaided or transferred to the clinical setting. These explanations are didactic hypotheses, not causal mechanisms demonstrated by synthesis (16, 18, 22).

The effect on cognitive load is another possible explanation. Integrated information within the field of action could facilitate orientation and coordination; a complex interface could also add demands unrelated to the learning objective. These possibilities require valid measures and designs that separate content, exposure time, feedback, and device. The summary by Xiao et al. provides a clue in histology, but its limited documentary access prevents verification of the mechanisms and does not allow generalization to other interventions (20). Tang et al. already pointed out limitations in the quality, validation, and feasibility of research on AR in 2020 (32). This synthesis adds a contrast between immediate acquisition, retention, and transfer in a more recent corpus of undergraduate education. Its contribution is interpretive and should be considered within the limits of its research and verification. Future evaluations should select instruments consistent with the competency, justify their measurement properties, report confidence intervals, and schedule transfer and follow-up tasks.

The institutional implications require distinguishing between educational outcomes and implementation outcomes. Non-inferiority in two POCUS measurements does not prove savings, sustainability, or general teacher replacement (28). The reviews by Betancourt Masri et al. and Zelada Espejo et al., published in this journal, provide context on technology teaching and three-dimensional models, but do not establish the cost-effectiveness of AR/RM (33-34). The adoption

decision requires considering costs for acquisition, maintenance, licensing, training, accessibility, and teaching time, along with sustained educational outcomes.

The incorporation of digital resources must respond to specific educational objectives and contexts. A quasi-experimental study in Medical Biochemistry reported academic improvements with intelligent learning and thinking routines (35), while a reflection on virtuality in Latin America highlighted opportunities for access and flexibility (36). None of these studies evaluated AR/RM, so they are cited as pedagogical background and not as evidence of its effectiveness.

The literature coverage is not exhaustive: it focused on PubMed/MEDLINE and references from previous reviews, without documented equivalent searches in educational and commercial bibliographic databases. Daily dates, exports, and complete screening decisions are not retained, and independent duplicate review or formal consensus cannot be demonstrated. RoB 2/ROBINS-I assessments for each outcome and a GRADE assessment of certainty were also not completed. These limitations preclude treating the synthesis as a comprehensive systematic review and restrict any curriculum recommendations. Evidence from abstracts or mixed populations is identified as supplementary. Heterogeneity, limited follow-up, small sample sizes, and variable denominators further limit the precision and transferability of inferences.

4.1. Implications for curriculum decision-making

Table 3 presents a list of criteria that can be used for curriculum decision-making, planning, and evaluating implementations. It is an interpretive framework, not a set of proven effects or a graded recommendation.

Table 3. Criteria for curriculum decision.

Curriculum question	Recommended criteria	Quality indicator	Risk avoided
What skill is the aim to develop?	Define knowledge, skill, or observable performance.	Explicit objective and standard.	Confusing technological exposure with learning.
What exactly does AR/MR offer?	Use it when adding relevant visualization, contextual guidance, or interaction.	Functional justification of the resource.	Technocentrism.
How will it be integrated?	Combine with active practice, feedback, and teacher guidance when needed.	Reproducible sequence.	Isolated or demonstrative use.
How will it be evaluated?	Prioritize objective testing, OSCE/OSPE, accuracy, errors, and time.	Instruments with evidence of validity.	Confusing satisfaction with competence.
Is it maintained and transferred?	Include deferred assessment and new or clinical tasks.	Retention and transfer.	Overestimating the immediate post-test.
Is it sustainable?	Analyze devices, technical support, teaching time, and scalability.	Total cost and frequency of use.	Expensive and underused acquisition.

5. Conclusions

- The studies reviewed suggest that AR/MR can benefit specific educational tasks, especially when it incorporates relevant spatial information or procedural guidance. Knowledge and follow-up results are heterogeneous; the absence of significant differences does not

demonstrate equivalence. Evidence of clinical transfer is limited, and the non-inferiority observed in point-of-care studies should be interpreted in light of their outcomes, margins, and analytical exclusions.

- Curriculum integration can be explored through local assessments with explicit objectives, appropriate comparators, and measurements of performance, retention, transfer, and costs. These proposals guide research and implementation; they do not represent proven institutional benefits. Satisfaction and confidence are complementary outcomes. Conclusions should be limited to the tasks and conditions studied, avoiding the interpretation of technological novelty as sufficient evidence of educational effectiveness.

Authorship contributions: Conceptualization: LJMM and YBG. Methodology: YBG and LBVC. Search and selection of studies: LJMM, YBG, and LBVC. Data extraction and curation: LJMM and LBVC. Analysis, synthesis, and interpretation of evidence: LJMM, YBG, and LBVC. Drafting of the original manuscript: LJMM and YBG. Critical review and editing of the manuscript: LJMM, YBG, and LBVC. Supervision: YBG.

Financing: The research did not receive funding.

Conflicts of interest: The authors declare no conflicts of interest related to this work.

Ethical considerations: The review used exclusively published information and did not include identifiable individual data; therefore, it did not require informed consent or approval from an ethics committee.

Statement on the use of artificial intelligence : In the literature review phase, Perplexity AI was used to help identify and review potentially relevant publications. This tool does not replace the authors' scientific judgment nor will it generate primary data. The selection of studies, the final verification of sources and results, the interpretation of the evidence, and the approval of the manuscript are the responsibility of the authors.

Supplementary material: preserved bibliographic strategy, critical characterization of the reports examined and limits of inferences by outcome.

6. References

1. Barteit S, Lanfermann L, Bärnighausen T, et al. Augmented, Mixed, and Virtual Reality-Based Head-Mounted Devices for Medical Education: Systematic Review. *JMIR Serious Games*. **2021**, 9, e29080. <https://doi.org/10.2196/29080>
2. Dhar P, Rocks T, Samarasinghe RM, et al. Augmented reality in medical education: students' experiences and learning outcomes. *Med Educ Online*. **2021**, 26, 1953953. <https://doi.org/10.1080/10872981.2021.1953953>
3. Baashar Y, Alkawsi G, Wan Ahmad WNW, et al. Effectiveness of Using Augmented Reality for Training in the Medical Professions: Meta-analysis. *JMIR Serious Games*. **2022**, 10, e32715. <https://doi.org/10.2196/32715>
4. Zhang R, Jin X, Liu M, et al. The effectiveness of augmented reality/mixed reality in medical education: a meta-analysis. *BMC Med Educ*. **2025**, 25, 1586. <https://doi.org/10.1186/s12909-025-08166-8>
5. Williams A, Sun Z, Vaccarezza M. Comparison of augmented reality with other teaching methods in learning anatomy: A systematic review. *Clin Anat*. **2025**, 38, 168-185. <https://doi.org/10.1002/ca.24234>
6. García-Robles P, Cortés-Pérez I, Nieto-Escámez FA, et al. Immersive virtual reality and augmented reality in anatomy education: A systematic review and meta-analysis. *Anat Sci Educ*. **2024**, 17, 514-528. <https://doi.org/10.1002/ase.2397>
7. Moro C, Birt J, Stromberga Z, et al. Virtual and Augmented Reality Enhancements to Medical and Science Student Physiology and Anatomy Test Performance: A Systematic Review and Meta-Analysis. *Anat Sci Educ*. **2021**, 14, 368-376. <https://doi.org/10.1002/ase.2049>
8. Mahmud M, Sari DCR, Sari D, et al. The application of augmented reality for improving clinical skills: a scoping review. *Korean J Med Educ*. **2024**, 36, 65-79. <https://doi.org/10.3946/kjme.2024.285>

9. González ST, Ibeas C, Gravert I, et al. Characteristics of a Virtual Reality scenario for learning anatomy: A literature review. *Rev Esp Edu Med*. **2022**, 3, 81-92. <https://doi.org/10.6018/edumed.542861>
10. Ampuero IP, Riquelme-Vilches V, Vargas E, et al. Influence of Technological Tools on the Teaching of Morphology in Medical Education: A Systematic Review. *Rev Esp Edu Med*. **2024**, 5, 631071. <https://doi.org/10.6018/edumed.631071>
11. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ*. **2020**, 368, 16890. <https://doi.org/10.1136/bmj.l6890>
12. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomized trials. *BMJ*. **2019**, 366, 14898. <https://eprints.whiterose.ac.uk/id/eprint/150579/>
13. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. **2016**, 355, i4919. <https://doi.org/10.1136/bmj.i4919>
14. Weeks JK, Pakpoor J, Park BJ, et al. Harnessing Augmented Reality and CT to Teach First-Year Medical Students Head and Neck Anatomy. *Acad Radiol*. **2021**, 28, 871-876. <https://doi.org/10.1016/j.acra.2020.07.008>
15. Stojanovska M, Tingle G, Tan L, et al. Mixed Reality Anatomy Using Microsoft HoloLens and Cadaveric Dissection: A Comparative Effectiveness Study. *Med Sci Educ*. **2020**, 30, 173-178. <https://doi.org/10.1007/s40670-019-00834-x>
16. Pickering JD, Panagiotis A, Ntakakis G, et al. Assessing the difference in learning gain between a mixed reality application and drawing screencasts in neuroanatomy. *Anat Sci Educ*. **2022**, 15, 628-635. <https://doi.org/10.1002/ase.2113>
17. Zeedzen I, Karstens J, van den Hurk M, et al. Comparing the effectiveness of augmented reality and anatomical atlases in student preparation for neuroanatomy dissection. *Sci Rep*. **2024**, 14, 24939. <https://doi.org/10.1038/s41598-024-76379-w>
18. Zingg L, Häusler M, Hein J, et al. Using Hologram-Based Augmented Reality in Anatomy Learning: The TEACHANATOMY Randomized Trial. *Acad Med*. **2025**, 100, 695-704. <https://doi.org/10.1097/ACM.0000000000006012>
19. Starszak K, Karaś R, Bajor G, et al. Advancing anatomical education: comparative effectiveness of mixed reality and 3D printing technologies. *BMC Med Educ*. **2026**, 26, 734. <https://doi.org/10.1186/s12909-026-08959-5>
20. Xiao R, Chen X, Yang X, et al. Enhancing Histological Learning Through Augmented Reality: A Comparative Study of Traditional Microscopy and AR-Based Instruction. *Clin Anat*. **2026**, advance online publication. <https://doi.org/10.1002/ca.70139>
21. Schoeb DS, Schwarz J, Hein S, et al. Mixed reality for teaching catheter placement to medical students: a randomized single-blind, prospective trial. *BMC Med Educ*. **2020**, 20, 510. <https://doi.org/10.1186/s12909-020-02450-5>
22. Guha P, Lawson J, Minty I, et al. Can mixed reality technologies teach surgical skills better than traditional methods? A prospective randomized feasibility study. *BMC Med Educ*. **2023**, 23, 144. <https://doi.org/10.1186/s12909-023-04122-6>
23. Hadida Barzilai D, Tejman-Yarden S, Yogev D, et al. Augmented Reality-Guided Mastoidectomy Simulation: A Randomized Controlled Trial Assessing Surgical Proficiency. *Laryngoscope*. **2025**, 135, 894-900. <https://doi.org/10.1002/lary.31791>
24. Hayasaka T, Kawano K, Onodera Y, et al. Comparison of accuracy between augmented reality/mixed reality techniques and conventional techniques for epidural anesthesia using a practice phantom model kit. *BMC Anesthesiol*. **2023**, 23, 171. <https://doi.org/10.1186/s12871-023-02133-w>
25. Farshad NA, Kubik RA, Kolling C, et al. Learning how to perform ultrasound-guided interventions with and without augmented reality visualization: a randomized study. *Eur Radiol*. **2023**, 33, 2927-2934. <https://doi.org/10.1007/s00330-022-09220-5>

26. Sarkar K, Coote J, Abdul R, et al. Feasibility, Adoption, and Utility of Augmented Reality in Lumbar Puncture Simulation by Medical Students: A Pilot Study. *J Med Ext Real*. **2024**, 1, 191-202. <https://doi.org/10.1089/jmxr.2024.0001>
27. Felten R, Bigaut K, Wirth T, et al. Advancing medical training with augmented reality and haptic feedback simulator: outcomes of a randomized controlled trial on lumbar puncture. *BMC Med Educ*. **2025**, 25, 1231. <https://doi.org/10.1186/s12909-025-07536-6>
28. Mulder TA, Jacobs RC, Zuurmond MS, et al. Augmedicine: Ultrasound: Training in ultrasound using XR – a randomized controlled trial. *BMC Med Educ*. **2026**, 26, 778. <https://doi.org/10.1186/s12909-026-08897-2>
29. Noll C, von Jan U, Raap U, et al. Mobile Augmented Reality as a Feature for Self-Oriented, Blended Learning in Medicine: Randomized Controlled Trial. *JMIR Mhealth Uhealth*. **2017**, 5, e139. <https://doi.org/10.2196/mhealth.7943>
30. Malik R, Abbas JR, Jayarajah C, et al. Mixed Reality Enhanced Otolaryngology Case-Based Learning: A Randomized Educational Study. *Laryngoscope*. **2023**, 133, 1606-1613. <https://doi.org/10.1002/lary.30364>
31. d'Aiello AF, Cabitza F, Natali C, et al. The Effect of Holographic Heart Models and Mixed Reality for Anatomy Learning in Congenital Heart Disease: An Exploratory Study. *J Med Syst*. **2023**, 47, 64. <https://doi.org/10.1007/s10916-023-01959-8>
32. Tang KS, Cheng DL, Mi E, et al. Augmented reality in medical education: a systematic review. *Can Med Educ J*. **2020**, 11, e81-e96. <https://doi.org/10.36834/cmej.61705>
33. Betancourt Masri L, Bello Astorga W, Osses Alvarado I, et al. Influence of the use of new technologies in medical semiology education. Systematic review. *Rev Esp Edu Med*. **2024**, 5, 631521. <https://doi.org/10.6018/edumed.631521>
34. Zelada Espejo MI, Segura Beltrán F, Margas Cavieres F, et al. Impact of 3D printing on the learning of medical students: a systematic review. *Rev Esp Edu Med*. **2024**, 5, 626811. <https://doi.org/10.6018/edumed.626811>
35. Batista GY, Villavicencio Cedeño LB. Use of intelligent learning and thinking routines in the teaching of Medical Biochemistry. *Educ Med Super*. **2025**, 39, e4667. <https://ems.sld.cu/index.php/ems/article/view/4667>
36. Batista GY, Villavicencio Cedeño LB. Virtuality in medical education: opportunities for Latin America. *Investig Educ Med*. **2026**, 15(57), 132-133. <https://doi.org/10.22201/fm.20075057e.2026.57.25747>

