

Review

Evaluation of the Effectiveness of Serious Games on the Learning of Clinical Skills in Health Science Students: A Systematic Review

Khadija Aboukad ^{1,*}, Mohamed Amine Baba ²  and Hicham Nassik ^{1,3} 

¹ Research and Innovation Laboratory in Health Science, Faculty of Medicine and Pharmacy, Ibn Zohr University, BP 7519, Quartier Tilila, Agadir CP80060, Morocco; h.nassik@uiz.ac.ma

² Higher Institute of Nursing Professions and Health Techniques of Agadir, Ministry of Health and Social Protection, Agadir CP80000, Morocco; babamedamine2@gmail.com

³ Department of Anesthesiology and Reanimation, Faculty of Medicine and Pharmacy of Agadir, University Ibn Zohr, Agadir CP80000, Morocco

* Correspondence: khadijaaboukad18@gmail.com; Tel.: +212-665097472

Abstract

Purpose: To evaluate the effectiveness of serious games, including virtual reality-based interventions, in improving clinical skills acquisition among undergraduate and postgraduate health science students. **Methods:** This systematic review was prospectively registered in PROSPERO (CRD42024589035) and conducted in accordance with PRISMA 2020 guidelines. PubMed, Scopus, Web of Science, and ScienceDirect were searched from inception to 31 August 2025. Eligible studies examined serious games, simulation-based platforms, or immersive and non-immersive virtual reality interventions designed to support clinical skills development. Outcomes were classified using a predefined hierarchical framework aligned with Miller's pyramid, distinguishing performance-based clinical competence, clinical reasoning, and secondary educational outcomes. Owing to substantial heterogeneity in interventions, comparators, and assessment methods, a narrative synthesis was performed. **Results:** Thirteen studies involving 892 participants were included. Serious games and virtual reality-based interventions were associated with generally favorable outcomes for knowledge acquisition, self-efficacy, motivation, satisfaction, and anxiety reduction. Improvements in clinical reasoning were reported in several studies, and some studies demonstrated benefits in performance-based clinical competence, particularly in simulation and virtual reality settings. However, findings for objective performance-based outcomes were mixed, with some studies reporting no statistically significant between-group differences. Heterogeneity in outcome definitions and limited reporting of standardized effect sizes reduced cross-study comparability. **Conclusions:** Serious games, including virtual reality-based interventions, may serve as complementary, scenario-based learning strategies in health sciences education. The most consistent effects were observed for cognitive and learner-centered outcomes, whereas evidence for objective gains in performance-based clinical competence remains variable. Further high-quality studies using standardized outcome frameworks, validated performance-based assessments, effect sizes, confidence intervals, and longer follow-up are needed.

Keywords: clinical skills; serious games; education; medical; students; health occupations; virtual reality



Academic Editor: Hideki Kasuya

Received: 2 May 2026

Revised: 9 June 2026

Accepted: 10 June 2026

Published: 18 June 2026

Copyright: © 2026 by the authors.

Published by MDPI on behalf of the

Academic Society for International

Medical Education. Licensee MDPI,

Basel, Switzerland. This article is an

open access article distributed under

the terms and conditions of the

[Creative Commons Attribution](#)

[\(CC BY\)](#) license.

1. Introduction

The acquisition of clinical skills is a cornerstone of health science education, ensuring that future professionals are prepared to deliver safe and effective patient care in increasingly complex health systems [1]. Clinical skills can be understood as multidimensional abilities that require the integration of knowledge, clinical reasoning, procedural performance, communication, decision-making, and safety-related behaviors to perform clinical tasks effectively and provide safe patient care [2,3]. Conventional instructional approaches, such as lectures, supervised practice, and clinical placements, remain essential but face important limitations, including uneven clinical exposure, restricted training opportunities, variability in supervision, and institutional pressures that may hinder optimal skill acquisition [4]. To address these challenges, educators have increasingly adopted innovative pedagogical strategies, including serious games. Serious games are purpose-designed game-based interventions developed primarily for education, training, skill acquisition, or behavioral change rather than entertainment alone [5]. In health sciences education, serious games usually combine explicit learning objectives, clinical scenarios, interactive decision-making, feedback mechanisms, scoring or progression systems, and game-based motivational features. By engaging learners in simulated clinical situations, serious games provide safe and interactive environments in which students can practice decision-making, observe the consequences of their actions, repeat tasks, and receive feedback before exposure to real patients. These characteristics may promote active learning, knowledge retention, learner engagement, self-confidence, and selected dimensions of clinical competence [6].

Over the past decade, research on serious games in health professions education has expanded considerably. Several studies and reviews have reported positive educational outcomes, including improvements in knowledge acquisition, clinical skills, learner satisfaction, motivation, and self-confidence among learners using serious game-based interventions compared with traditional teaching approaches [7–9]. Similarly, Lee et al. (2024) demonstrate that digital serious games may improve knowledge acquisition, performance-related outcomes, and confidence among nursing students, suggesting that these interventions may contribute to specific dimensions of clinical competence when they include structured clinical scenarios, decision-making tasks, feedback, and objective assessment of learner performance [10].

Despite these encouraging findings, the evidence remains fragmented because studies differ substantially in populations, clinical domains, intervention formats, comparator conditions, and outcome measures. Importantly, knowledge-based and learner-centered outcomes are often reported alongside higher-order outcomes such as clinical reasoning and performance-based competence, although these outcomes represent different levels of learning. This creates a need for a synthesis that distinguishes cognitive, reasoning-related, and performance-based outcomes. Therefore, this systematic review aimed to assess the effectiveness of serious games, including virtual reality-based interventions, in enhancing clinical skills acquisition among health science students, while differentiating performance-based clinical competence, clinical reasoning, and secondary educational outcomes.

2. Materials and Methods

2.1. Protocol and Registration

This systematic review was prospectively registered in the International Prospective Register of Systematic Reviews (PROSPERO; registration number CRD42024589035). The review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines [11].

2.2. Eligibility Criteria

The eligibility criteria were defined according to the PICOS framework (Population, Intervention, Comparator, Outcomes, Study design).

2.3. Population

Studies were eligible if they included undergraduate or postgraduate students enrolled in health sciences programs (e.g., medicine, nursing, dentistry, pharmacy, physiotherapy, or related disciplines). Studies involving practicing professionals exclusively were excluded unless data for students were reported separately.

2.4. Intervention

Eligible interventions included serious games, immersive or non-immersive virtual reality (VR) applications, simulation games, and gamified educational platforms explicitly designed to train clinical competencies. For the purpose of this review, serious games were considered the overarching category of game-based educational interventions explicitly designed for learning or training rather than entertainment alone. VR-based interventions were included when they incorporated interactive clinical scenarios, learner decision-making, feedback, and explicit educational objectives related to clinical skills development. Simulation-based platforms were included when they provided structured, scenario-based clinical learning experiences with interactive or game-like components. Gamification-based interventions were included only when game elements, such as points, scoring, feedback, progression, or challenges, were embedded within a structured clinical learning activity and were used to support clinical skills, clinical reasoning, or related educational outcomes. Purely decorative gamification or reward systems without clinical scenario interaction were excluded.

To be considered for inclusion, interventions were required to incorporate core pedagogical and interactive features consistent with competency-based training. Specifically, they had to integrate structured clinical scenarios or case-based decision-making processes, promote active learner engagement through clinical decision-making tasks, and include feedback mechanisms provided either immediately or after task completion. In addition, eligible interventions could involve rule-based progression systems, scoring structures, or performance tracking components, and were expected to articulate explicit educational objectives focused on the development of clinical skills.

Both digital serious games, including desktop-based, web-based, mobile, three-dimensional, and immersive virtual reality applications, and structured non-digital serious games were considered. Examples of non-digital serious games included board games, card-based clinical decision-making games, tabletop simulation games, and structured role-play games, provided that they were designed around clinical scenarios, included explicit educational objectives, promoted learner interaction, and assessed clinical skills, clinical reasoning, or related educational outcomes.

2.5. Comparator

Eligible comparators included traditional lectures, e-learning modules, problem-based learning (PBL), high-fidelity simulation, self-directed learning materials, or alternative educational strategies.

Studies without a comparator group were included for descriptive synthesis if they reported pre–post changes aligned with the predefined outcome framework.

2.6. Outcomes

To ensure conceptual clarity and methodological consistency, outcomes were structured hierarchically according to an educational competency framework aligned with Miller's pyramid [12]. This predefined framework was used to distinguish between different levels of clinical learning and competence. Performance-based clinical competence was considered the highest-level outcome and referred to objective assessments of students' ability to perform clinical or procedural tasks in simulated or structured settings, such as OSCE scores, structured performance checklists, simulation-based assessments, or observer-rated clinical tasks. Clinical reasoning was considered an intermediate higher-order cognitive outcome and included structured assessments of diagnostic reasoning, case-based decision-making, clinical problem-solving, key-feature examinations, or algorithm-based performance indicators. Knowledge acquisition and learner-centered outcomes, including self-efficacy, motivation, satisfaction, anxiety, engagement, usability, and learner experience, were classified as secondary outcomes. This classification was used to avoid interpreting knowledge-only or self-reported outcomes as direct evidence of performance-based clinical competence.

2.7. Primary Outcomes

The primary outcomes of this review were performance-based clinical competence and clinical reasoning. Performance-based clinical competence referred to the objective evaluation of students' ability to perform procedural or clinical tasks using standardized assessment methods. These measures included Objective Structured Clinical Examination (OSCE) scores, structured performance checklists, simulation-based performance assessments, observer-rated clinical tasks, and other objective indicators reflecting procedural accuracy or successful task completion.

Clinical reasoning was considered a primary cognitive competence outcome. It referred to the objective or structured evaluation of diagnostic reasoning, clinical decision-making, case-based problem-solving, and higher-order cognitive processes. These measures included key-feature examinations, standardized case-based scoring systems, algorithm-based log-file performance metrics, and other structured clinical reasoning assessments.

2.8. Secondary Outcomes

Secondary outcomes included knowledge acquisition and learner-centered educational outcomes. Knowledge acquisition was commonly evaluated using multiple-choice questions or written knowledge tests. Learner-centered outcomes included self-efficacy, motivation, satisfaction, anxiety, engagement, usability, and learner experience. Studies reporting knowledge-only outcomes were included and analyzed separately within the secondary outcomes category. In accordance with Miller's framework of clinical competence, these outcomes were interpreted as lower-level educational outcomes and were not considered direct indicators of performance-based clinical competence.

2.9. Study Design

Eligible study designs included randomized controlled trials (RCTs), cluster RCTs, quasi-experimental controlled studies, and prospective cohort studies. Feasibility studies and uncontrolled pre-post studies were included only in the descriptive synthesis when they reported educational outcomes consistent with the predefined outcome framework. Their inclusion was justified by the emerging and heterogeneous nature of serious game- and virtual reality-based interventions in health sciences education, where early-stage studies may provide relevant information on intervention design, acceptability, implementation, and preliminary educational outcomes. However, these studies were not interpreted as

providing strong evidence of comparative effectiveness because they lacked a control group or were not designed to estimate between-group effects.

Conference abstracts, editorials, case reports, and narrative reviews were excluded.

2.10. Information Sources and Search Strategy

A comprehensive literature search was conducted in PubMed, Scopus, Web of Science, and ScienceDirect from database inception to 31 August 2025.

The search was restricted to studies published in English, as this was the language in which the review team could reliably assess full texts and extract data.

The search strategy combined controlled vocabulary, when available, and free-text terms related to serious games, virtual reality, gamification, simulation-based learning, clinical skills, clinical competence, health science students, medical education, nursing education, and health professions education. Boolean operators, truncation, phrase searching, and database-specific field tags were applied where appropriate. Reference lists of included studies were also manually screened to identify additional eligible articles. The complete reproducible search strings for each database are provided in Supplementary Material S1.

2.11. Study Selection

All records were imported into Catchii software for screening and management.

Two reviewers independently screened titles and abstracts for eligibility. Full texts of potentially relevant studies were assessed against predefined inclusion and exclusion criteria. Disagreements were resolved through discussion or consultation with a third reviewer.

A PRISMA flow diagram summarizes the study selection process.

Data Extraction

Data were independently extracted by two reviewers using a standardized data extraction form. Extracted information included study characteristics (author, year of publication, country, and study setting), study design and methodological features (randomization procedures, allocation methods, and blinding), as well as sample size and participant characteristics. Detailed information on the intervention was collected, including the type of serious game, level of immersion, duration of exposure, delivery platform, and feedback mechanisms. Comparator characteristics were also recorded where applicable. Outcome measures were documented, including the type of outcomes assessed, the validity of measurement instruments, and their classification as performance-based or cognitive competence outcomes. Statistical analysis methods and the main study findings, including reported means, standard deviations, and effect sizes when available, were also extracted. When required, corresponding authors were contacted to obtain missing or additional data.

2.12. Risk-of-Bias and Methodological Quality Assessment

The methodological quality of randomized controlled trials was assessed using the Cochrane Risk of Bias 2 (RoB 2) tool, which evaluates bias arising from the randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result. Non-randomized studies were evaluated using the ROBINS-I tool, which assesses bias due to confounding, selection of participants, classification of interventions, deviations from intended interventions, missing data, measurement of outcomes, and selection of the reported result.

Risk-of-bias assessments were conducted independently by two reviewers. Inter-rater agreement before consensus was calculated using Cohen's kappa coefficient. Discrepancies between reviewers were then resolved through discussion and consensus; when disagreement persisted, a third reviewer was consulted. Domain-level judgments and overall risk-of-bias assessments were recorded for each included study.

Due to the anticipated heterogeneity in intervention characteristics, study designs, comparator groups, and outcome measures among the included studies, a narrative synthesis approach was adopted.

3. Results

3.1. Study Characteristics

A total of 13 studies met the eligibility criteria and were included in the final synthesis, with publication dates ranging from 2017 to 2024 [13–25] (Figure 1). In terms of study design, six studies were randomized controlled trials, including cluster and multi-arm randomized trials. Four studies used quasi-experimental designs, one study was a prospective comparative study, one was a feasibility mixed-methods study, and one study involved the development and evaluation of an educational intervention with a naturalistic comparison.

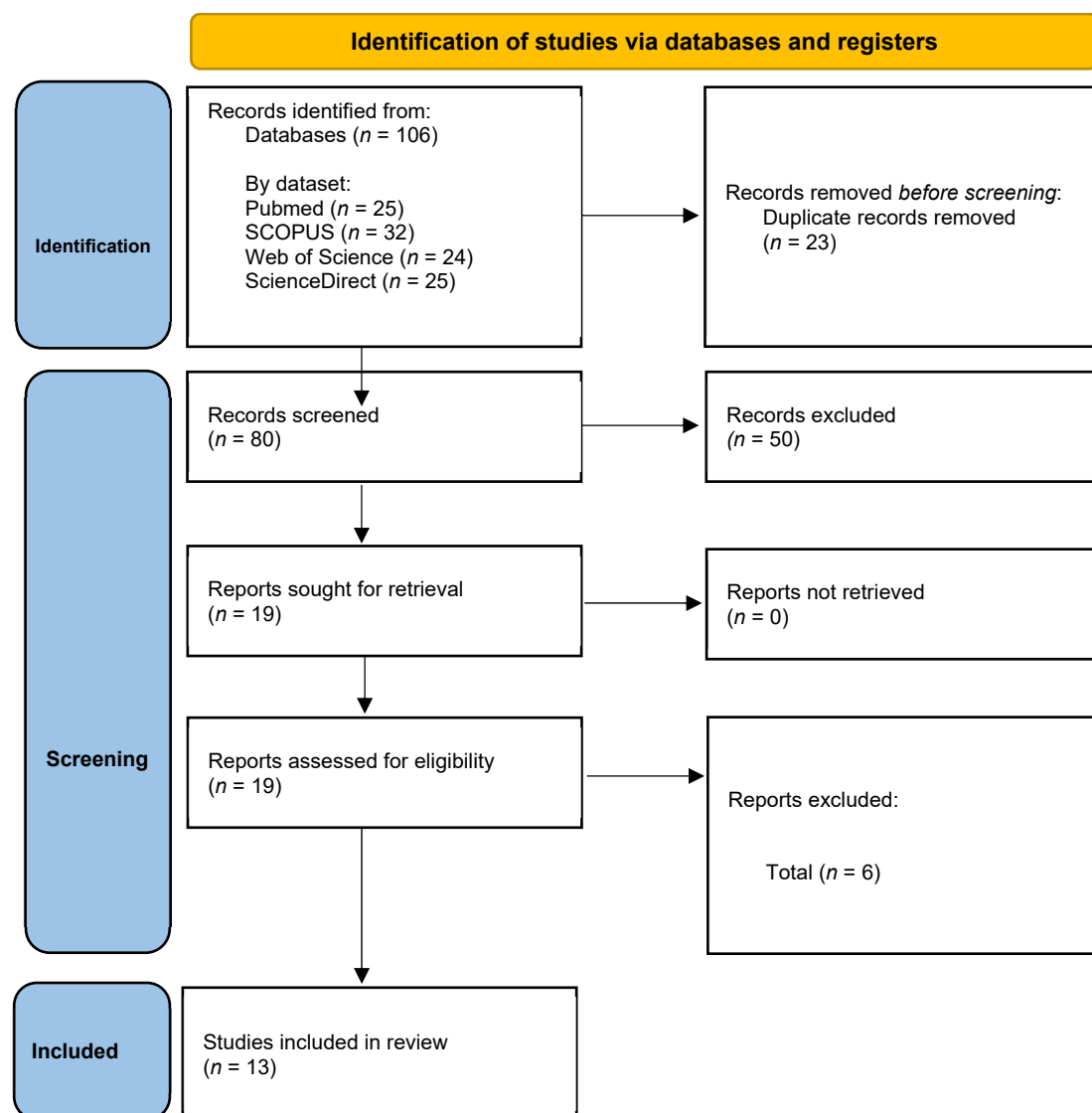


Figure 1. PRISMA Flow Diagram.

The studies were conducted across multiple geographic settings, including Singapore (n = 1) [23], France (n = 1) [16], Germany (n = 2) [20,21], Switzerland (n = 1) [15], Iran (n = 2) [17,19], Taiwan (n = 1) [24], the United Kingdom (n = 1) [13], South Korea (n = 1) [25], Finland (n = 1) [18], New Zealand (n = 1) [14] and Italy (n = 1) [22] (Table 1).

Table 1. Characteristics of the 13 Included Studies.

Study (Author, Year)	Country/Discipline	Study Design	Sample	Intervention (Type)	Comparator	Reported Outcomes (According to Framework)
[23]	Singapore/Nursing	Cluster Randomized Controlled Trial	103 second-year nursing students	Serious game: “Blood Transfusion”	Waitlist control	Performance-based clinical competence (simulation), Knowledge, Self-efficacy, Motivation, Satisfaction
[16]	France/Medicine	Randomized Controlled Trial	82 medical students (79 analyzed)	Serious game: “Staying Alive” (CPR pre-training)	Online lecture (PowerPoint)	Performance-based clinical competence (time to reach minimum passing score in simulation)
[16]	Germany/Medicine	Prospective Comparative Study	EMERGE group vs. PBL group (as reported in article)	Serious game: “EMERGE” (clinical reasoning)	Problem-Based Learning (small groups)	Clinical reasoning (key-feature examination), In-game performance metrics
[14]	New Zealand (Auckland)/Medicine	Pilot Mixed-Methods Study	46 students (Game: 27; PDF: 19)	3D Serious game: “Metaphoria” (history taking)	PDF-based learning material	Performance-based clinical competence (OSCE), Satisfaction, Other secondary outcomes
[21]	Germany/Medicine	Prospective Randomized Study (Three Arms)	61 analyzed (A = 25; B = 16; AB = 20)	Serious game (clinical reasoning and transfer training)	Variations in case exposure (according to study arms)	Clinical reasoning (transfer to comparable clinical problems)
[15]	Switzerland/Medicine	Pilot Randomized Controlled Trial	Medical students (sample reported in study)	Immersive VR COVID-19 diagnostic training (PPE, swab procedures)	Traditional non-VR training	Performance-based clinical competence (skills test in simulated scenario), Satisfaction, Usability
[19]	Iran/Medicine	Randomized Trial	50 medical students	VR-based serious game (approach to coma)	Traditional lecture	Knowledge acquisition (primary), Other secondary outcomes
[24]	Taiwan/Dentistry	Development and Evaluation Study	Application stage: 34 dental students	Serious game: “Virtual Dental Clinic (VDC)”	Naturalistic comparison (users vs. non-users)	Clinical reasoning (validity indicators), Satisfaction, Usability

Table 1. Cont.

Study (Author, Year)	Country/Discipline	Study Design	Sample	Intervention (Type)	Comparator	Reported Outcomes (According to Framework)
[13]	United Kingdom/Nursing	Feasibility Mixed-Methods Study	Stage 1: 19 nursing students	VR sepsis serious game (adjunct training)	No comparator (pre–post with qualitative component)	Self-efficacy and Anxiety (NASC-CDM scale)
[25]	South Korea/Nursing	Quasi-experimental Study (Three Groups)	VR: 29; Simulation: 28; Control: 26	Immersive VR gamification (NRP, ARCS-based design)	High-fidelity simulation + online lecture; Online lecture only	Knowledge, Motivation, Other secondary outcomes (problem-solving ability, self-confidence, anxiety)
[17]	Iran/Nursing	Multi-arm Randomized Controlled Trial	166 nursing students	Flipped classroom, Gamification, Combined intervention	Lecture-based or alternative arms	Performance-based clinical competence, Knowledge, Self-efficacy, Satisfaction
[18]	Finland/Nursing	Quasi-experimental (Pre–Post with Control)	276 students (Experimental: 140; Control: 136)	3D Unity-based simulation game	Self-study theoretical material	Knowledge acquisition (SNK test)
[22]	Italy/Nursing education (mental health context)	Quasi-experimental pre–post study	48 undergraduate nursing students (29 second-year, 19 third-year)	Gamification-based intervention: Serious Game (SG) teaching the SBAR (Situation, Background, Assessment, Recommendation) structured handover framework through psychiatric care scenarios	Pre-intervention performance (pre-test) using the same SBAR classification test	Primary outcome: significant improvement in SBAR application scores after the intervention (mean score increased from 13 to 16.2; $p < 0.001$; Cohen's $d = 0.78$). Secondary outcomes: positive student experience measured with the Player Experience Inventory (PXI) across domains such as Meaning, Curiosity, Progress Feedback, Audiovisual Appeal, Challenge, Ease of Control, Clarity of Goals, and Enjoyment

Across all included studies, the total sample comprised 892 participants. Individual study sample sizes ranged from 11 to 276 students. Most studies included undergraduate students in nursing ($n = 6$) [13,14,17,18,22,23] or medicine ($n = 6$) [15,16,19–21,25], and one study involved dental students [24].

3.2. Intervention Characteristics

All studies evaluated serious games or virtual reality-based educational interventions. Ten studies used digital serious games or virtual simulations delivered via desktop or web-based platforms [14–19,21–24]. Three studies employed immersive virtual reality systems [13,20,25].

Interventions were designed to simulate clinical scenarios, decision-making processes, communication tasks, or procedural tasks. Most studies reported the use of structured cases, interactive elements, and scoring or feedback mechanisms.

3.3. Comparator Characteristics

Comparator conditions varied substantially across the included studies. Lecture-based or conventional teaching was used in studies evaluating gamification, flipped classroom, or virtual reality-based serious gaming approaches, including Ghafouri et al. [17] and Mansoori et al. [19]. Online, text-based, or self-directed learning materials were used as comparators in several studies, including Drummond et al. [16], Tan et al. [23], and Alyami et al. [14]. Problem-based learning was used as the comparator in the study by Middeke et al. [20], whereas Birrenbach et al. [15] compared immersive VR training with established traditional training based on video and written instructions. Koivisto et al. [18] used self-study materials as the comparator. Yang and Oh [25] used multiple comparator conditions, including high-fidelity simulation and other educational strategies. Some studies did not include a conventional randomized comparator group or used self-selected exposure or pre–post designs, including Parozzi et al. [22] and Wu et al. [24].

3.4. Outcomes Reported

According to the predefined hierarchical framework, performance-based or performance-related clinical competence outcomes were reported in six studies [14–17,22,23]. Five studies assessed objective clinical or procedural performance using randomized or comparative designs, including simulation-based performance scores, OSCE scores, structured checklists, time-to-performance thresholds, or objective skills tests. One study, Parozzi et al., assessed performance-related clinical communication competence using a structured SBAR application test in an uncontrolled pre–post design and was therefore interpreted descriptively rather than as comparative evidence of effectiveness [22].

Clinical reasoning outcomes were reported in studies using structured assessment tools, including key-feature examinations, structured reasoning assessments, in-game performance metrics, or case-based decision-making indicators [17,20,21,24]. Ghafouri et al. was included in both the performance-based clinical competence and clinical reasoning categories because the study reported distinct outcome measures: students' performance in client health assessment and scores on 10 key-feature questions assessing the application of health assessment knowledge in clinical scenarios.

Knowledge acquisition outcomes were reported in studies using structured knowledge tests, MCQ-based assessments, or written examinations [17–19,23,25].

Learner-centered outcomes, including self-efficacy, motivation, satisfaction, anxiety, usability, engagement, and learner experience, were classified as secondary outcomes and analyzed separately from performance-based clinical competence [13–15,22]. Outcome assessment timing was primarily immediate post-intervention, with limited reporting of follow-up assessments beyond the short term (Table 2).

Table 2. Intervention Characteristics and Outcome Measures of Included Studies.

Study	Intervention Format	Level of Immersion	Clinical Scenario Focus	Duration/Exposure	Assessment Tools Used	Outcome Category (Framework)
[23]	Digital serious game	Non-immersive	Blood transfusion procedure	Single-session training	Simulation performance checklist; Knowledge test	PC; K
[16]	Digital serious game	Non-immersive	Basic life support (CPR)	Pre-training module	Simulation-based time-to-performance (MPS)	PC
[20]	Digital serious game (EMERGE)	Non-immersive	Emergency medicine cases	Repeated case-based sessions	Key-feature examination; In-game scoring metrics	CR
[14]	3D digital serious game (Metaphoria)	Non-immersive 3D	Clinical history taking	40 min intervention	OSCE; Satisfaction scale	PC; Secondary
[21]	Digital serious game	Non-immersive	Clinical reasoning transfer tasks	Multi-session exposure	Structured reasoning assessment (transfer tasks)	CR
[15]	Immersive VR simulation	Fully immersive VR	COVID-19 diagnostics (PPE, swab procedures)	Simulation session	Objective skills test in simulated environment	PC
[19]	VR-based serious game	Immersive VR	Approach to coma	Single session	Knowledge test; Usability scale	K
[24]	Digital serious game (VDC)	Non-immersive	Virtual dental clinical cases	Application phase	Structured reasoning indicators; Validity metrics	CR
[13]	Immersive VR serious game	Fully immersive VR	Sepsis recognition and management	Single-session feasibility	NASC-CDM (self-efficacy / anxiety)	Secondary
[25]	Immersive VR gamification	Fully immersive VR	Neonatal resuscitation (NRP)	Multi-session	Knowledge test; Motivation scale; Problem-solving scale	K; Secondary
[17]	Gamification/Flipped/Combined	Non-immersive	Nursing clinical performance	Multi-arm structured exposure	Performance checklist; Knowledge test; Self-efficacy scale	PC; K; Secondary
[18]	3D simulation game (Unity)	Non-immersive 3D	Nursing knowledge integration	Multi-session	SNK knowledge test	K
[22]	Serious game (interactive quiz scenarios)	Low–moderate	Psychiatric nursing handover (SBAR)	45 min (single session)	SBAR pre–post test; Player Experience Inventory (PXL)	Improved SBAR performance; positive learner experience

PC: Performance-based Clinical Competence, CR: Clinical Reasoning, K: Knowledge, SEC: Secondary Educational Outcomes.

3.5. Intervention Characteristics and Outcome Measures

Table 2 summarizes the characteristics of the interventions and outcome measures reported across the 13 included studies.

3.6. Intervention Format and Immersion Level

Of the 13 studies, ten evaluated digital serious games delivered via desktop, web-based, platform-based, or 3D computer formats [14–19,21–24]. Three studies employed fully immersive virtual reality (VR) systems [13,20,25] (Table 2).

Regarding immersion level, five interventions were categorized as immersive VR experiences [13,14,18,21,25], while the remaining digital interventions were non-immersive or semi-immersive 3D environments.

3.7. Clinical Scenario Focus

All interventions were designed around structured clinical scenarios. The clinical domains included blood transfusion procedures, cardiopulmonary resuscitation, emergency medicine case management, clinical history taking, COVID-19 diagnostics, coma assessment, sepsis management, neonatal resuscitation, dental clinical reasoning, and structured nursing handover in mental health [13–25].

Several studies explicitly reported case-based decision-making processes embedded within the game structure. Most interventions incorporated scoring systems, performance tracking, or structured progression through clinical cases.

3.8. Duration and Exposure

Intervention duration and exposure also varied substantially. Some studies used brief or single-session serious game or VR activities [13–15,18,19,23], whereas others used preparatory serious game exposure before simulation [16], repeated case-based EMERGE training [20,21], short pre–post gamified activities [22], or self-directed application during clinical training [24]. This variability limited direct comparability across studies and supported the use of narrative synthesis.

3.9. Outcome Assessment Tools

Outcome measures were classified according to the predefined hierarchical framework [12–24].

Performance-based clinical competence outcomes were assessed in five studies [14–16,22,23] using objective structured clinical examinations (OSCEs), simulation-based performance assessments, or structured performance checklists.

Clinical reasoning outcomes were reported in four studies [20,21,24,25] using key-feature examinations, structured reasoning assessments, or validated in-game performance indicators.

Knowledge acquisition was reported in seven studies [14,17,18,22–25] and was measured using structured knowledge tests or standardized written assessments.

Secondary outcomes, including self-efficacy, motivation, satisfaction, anxiety, usability, and learner experience, were reported in eight studies [13–15,17,22–25] using validated scales or structured questionnaires.

Alignment with the Predefined Framework

All included studies reported at least one outcome measure that aligned with the predefined outcome framework. Studies reporting knowledge-only outcomes were categorized under secondary outcomes. Studies without comparator groups were included in the descriptive synthesis and reported pre–post changes within the predefined framework.

3.10. Risk-of-Bias Assessment

The risk of bias was assessed using the Cochrane Risk of Bias 2 (RoB 2) tool for randomized studies and the ROBINS-I tool for non-randomized studies. Domain-level judgments for randomized studies assessed using RoB 2 are presented in Table 3, while domain-level judgments for non-randomized studies assessed using ROBINS-I are presented in Figure 2. Inter-rater agreement for the risk-of-bias assessment before consensus was almost perfect, with a Cohen’s kappa of 0.82.



Figure 2. Risk-of-bias assessment of non-randomized studies using the ROBINS-I traffic-light plot. Domain-level judgments are presented for Middeke et al. [19], Alyami et al. [13], Wu et al. [23], Adhikari et al. [12], Yang and Oh [24], Koivisto et al. [17], and Parozzi et al. [21]. Green indicates low risk, yellow indicates moderate risk, and red indicates serious risk.

3.11. Randomized Controlled Trials

Six studies employed randomized designs, including cluster, pilot, prospective randomized, and multi-arm randomized trials [15–17,19,21,23]. Across these studies, the randomization process was generally described; however, details regarding allocation concealment were not consistently reported. Blinding of participants and educators was generally not feasible because of the nature of educational interventions. Outcome assessor blinding was explicitly reported in only a limited number of studies.

The main sources of concern in randomized studies were related to incomplete reporting of allocation concealment, potential deviations from intended interventions, and limited reporting of blinding procedures. Missing outcome data were generally addressed adequately, with attrition reported or accounted for in most trials. No clear evidence of selective outcome reporting was identified based on comparison between outcomes described in the Methods sections and those reported in the Results sections. Overall, randomized studies were judged as having either low risk of bias in individual domains or some concerns, primarily because of limitations related to allocation concealment, deviations from intended interventions, and outcome measurement procedures.

Table 3. Risk-of-Bias Assessment of Included Studies.

Study	Bias Arising from the Randomization Process	Bias Due to Deviations from Intended Interventions	Bias Due to Missing Outcome Data	Bias in Measurement of the Outcome	Bias in Selection of the Reported Result	Overall Risk of Bias
[23]	Some concerns	Low risk	Low risk	Low risk	Low risk	Some concerns
[16]	Low risk	Some concerns	Low risk	Some concerns	Low risk	Some concerns
[21]	Some concerns	Some concerns	Low risk	Some concerns	Low risk	Some concerns
[15]	Some concerns	Some concerns	Low risk	Some concerns	Low risk	Some concerns
[19]	Some concerns	Some concerns	Low risk	Some concerns	Low risk	Some concerns
[17]	Some concerns	Some concerns	Low risk	Some concerns	Low risk	Some concerns

Non-Randomized and Quasi-Experimental Studies

Seven studies used non-randomized designs, including prospective comparative studies, quasi-experimental pre–post studies, feasibility studies, and development-and-evaluation studies [13,14,18,20,22,24,25]. According to ROBINS-I criteria, the overall risk of bias in these studies was generally judged as moderate, except for one uncontrolled pre–post study, which presented serious concerns related to confounding because the absence of a control group limited causal inference.

The main sources of bias in non-randomized studies were related to the absence of randomization, potential baseline confounding, convenience sampling, lack of control groups in some studies, and limited reporting of outcome assessor blinding. Outcome assessment procedures were variably reported across studies. In uncontrolled pre–post designs, observed post-intervention changes could not be clearly separated from testing effects, maturation, or other contextual factors. Attrition was reported in most studies, and missing data were addressed where applicable. No clear evidence of selective outcome reporting was identified based on the available descriptions. Therefore, findings from non-randomized and uncontrolled pre–post studies were interpreted cautiously and primarily as descriptive evidence rather than strong evidence of comparative effectiveness.

3.12. Quantitative Results for Primary Outcomes

Table 4 presents the quantitative findings for primary outcomes, including performance-based or performance-related clinical competence and clinical reasoning. Where available, effect estimates, standardized effect sizes, confidence intervals, and *p*-values were extracted from the original studies or calculated from available summary data. When standardized effect sizes or confidence intervals were not reported in the original studies and could not be calculated reliably from the available data, they were indicated as not reported. Given the heterogeneity of study designs, assessment tools, and statistical reporting, findings were interpreted by considering the direction, magnitude, and precision of effects where possible, rather than statistical significance alone.

3.13. Performance-Based or Performance-Related Clinical Competence

Performance-based or performance-related clinical competence outcomes were reported in six studies. Five studies assessed objective clinical or procedural performance using randomized or comparative designs, whereas one uncontrolled pre–post study was interpreted descriptively.

Tan et al. assessed simulation-based clinical performance using a checklist-based performance score. The intervention group achieved a mean score of 24.91 (SD 5.04) compared with 22.89 (SD 5.14) in the control group, corresponding to a mean difference of +2.02 points and an approximate Cohen's *d* of 0.40; however, the difference was not statistically significant (*p* = 0.105) [23]. Drummond et al. assessed performance using the time required to reach a minimum passing score during CPR simulation-based mastery learning. The serious game group reached the minimum passing score in a median time of 20.5 min (IQR 15.8–30.3), compared with 23 min (IQR 15–32) in the control group, corresponding to a median difference of –2.5 min (*p* = 0.51). The median number of attempts was also similar between groups [16].

Table 4. Quantitative Results for Primary Outcomes: Performance-Based or Performance-Related Clinical Competence and Clinical Reasoning.

Study	Outcome Type	Assessment Tool	n	Intervention Result	Comparator/Pre-Intervention Result	Effect Estimate	95% CI	p-Value
[23]	PC	Simulation-based performance score/checklist-based clinical performance	57/46	Mean 24.91 (SD 5.04)	Mean 22.89 (SD 5.14)	MD = +2.02; Cohen's d $\approx$ 0.40	NR	0.105
[16]	PC	Time to reach minimum passing score during CPR simulation-based mastery learning	40/39 assessed for primary outcome	Median 20.5 min (IQR 15.8–30.3); median attempts 3 (IQR 2–5)	Median 23 min (IQR 15–32); median attempts 4 (IQR 2–5)	Median difference = −2.5 min; standardized effect size NR	NR	0.51 for training time; 0.52 for number of attempts
[20]	CR	Key-feature examination and final EMERGE session	78/34	Key-feature total score: median 62.5% (IQR 17.7)	Key-feature total score: median 54.2% (IQR 21.9)	Median difference = +8.3 percentage points; standardized effect size NR	NR	0.015
[14]	PC	OSCE history-taking station	27/19	Mean 15.04 (SD 4.28)	Mean 14.33 (SD 4.52)	MD = +0.71; Cohen's d $\approx$ 0.16	NR	0.60
[21]	CR	Final EMERGE session; aggregate clinical reasoning scores across four comparable virtual patient cases	Total analyzed $n = 61$; exposed vs. unexposed comparisons	-NSTEMI: 67.7% (SD 11.3); Asthma: 65.0% (SD 15.9); Pancreatitis: 69.2% (SD 11.6); GI hemorrhage: 61.8% (SD 12.5)	NSTEMI: 58.8% (SD 13.8); Asthma: 58.8% (SD 14.0); Pancreatitis: 65.2% (SD 19.8); GI hemorrhage: 62.8% (SD 9.6)	SMD $\approx$ 0.74 for NSTEMI; 0.41 for Asthma; 0.28 for Pancreatitis; −0.09 for GI hemorrhage	NR	NSTEMI $p = 0.019$; asthma $p = 0.015$; pancreatitis $p = 0.7454$; GI hemorrhage $p = 0.958$
[15]	PC	Nasopharyngeal swab performance checklist, 17 items	15/14	Post-test 1: median 14/17 (IQR 13–15)	Post-test 1: median 12/17 (IQR 11–14)	Median difference = +2 points; standardized effect size NR	NR	0.03

Table 4. Cont.

Study	Outcome Type	Assessment Tool	n	Intervention Result	Comparator/Pre-Intervention Result	Effect Estimate	95% CI	p-Value
[17]	CR/performance-related health assessment competence	Key-feature questions assessing client health assessment	166 total	Post-intervention overall mean = 17.98 (SD 1.25)	Pre-intervention overall mean = 12.90 (SD 1.03)	Mean change = +5.08; standardized effect size NR	NR	0.001 for overall pre-post comparison; ANOVA $p < 0.001$ for between-group post-intervention scores
[22]	PC-related clinical communication competence	SBAR structured handover classification test, 26 items	48/N/A	Post-test mean 16.2 (SD 4.28); median 16.0	Pre-test mean 13.0 (SD 3.74); median 13.5	Mean difference = +3.2 points; Cohen's $d = 0.78$; rank-biserial correlation = 0.414	Approx. 95% CI for MD = 1.57 to 4.83	<0.001
[24]	CR/performance-related dental clinical competence	National qualification test assessing clinical reasoning, treatment planning, and instrument selection; clerkship specialty performance scores	34 total; VDC use: once $n = 9$, ≥ 2 times $n = 12$; no VDC use $n = 13$	Qualification test: VDC once 80.00 (SD 5.59); VDC ≥ 2 times 79.58 (SD 6.90)	No VDC use: 70.00 (SD 13.54)	Qualification test: MD any VDC use vs. no use = +9.76; Cohen's $d \approx 1.01$. Conservative dentistry: MD = +1.65; Cohen's $d \approx 0.94$	NR	Qualification test $p = 0.029$; Periodontics $p = 0.037$; Conservative dentistry $p = 0.040$

PC, performance-based clinical competence; CR, clinical reasoning; OSCE, Objective Structured Clinical Examination; CPR, cardiopulmonary resuscitation; SBAR, Situation, Background, Assessment, Recommendation; MD, mean difference; SMD, standardized mean difference; IQR, interquartile range; CI, confidence interval; NR, not reported in the original study or not calculable from the available summary data. Parozzi et al. used an uncontrolled pre-post design without a control group; therefore, its findings were interpreted descriptively and not as comparative evidence of effectiveness. Ghafouri et al. was classified under clinical reasoning because the study used key-feature questions to assess students' application of health assessment knowledge in clinical scenarios. Parozzi et al. used an uncontrolled pre-post design without a control group and without individual-level pre-post matching. Therefore, its findings were interpreted descriptively and not as comparative evidence of effectiveness. The approximate 95% CI for the mean difference was calculated from the available summary data using the independent-samples framework reported in the original study. Ghafouri et al. was classified under clinical reasoning because the study used key-feature questions to assess students' application of health assessment knowledge in clinical scenarios. The study also contributed to the broader interpretation of performance-related health assessment competence.

Alyami et al. evaluated performance using an OSCE history-taking station. Mean OSCE scores were 15.04 (SD 4.28) in the serious game group and 14.33 (SD 4.52) in the comparator group, corresponding to a mean difference of +0.71 points and an approximate Cohen's *d* of 0.16 ($p = 0.60$) [14]. Birrenbach et al. assessed COVID-19-related procedural performance using a nasopharyngeal swab checklist. Immediately after training, the VR group achieved a higher median score than the control group, with 14/17 points (IQR 13–15) versus 12/17 points (IQR 11–14), corresponding to a median difference of +2 points ($p = 0.03$). At one-month follow-up, the difference was no longer significant [15].

Ghafouri et al. assessed performance-related health assessment competence and cognitive application using key-feature questions. Overall scores improved from a mean of 12.90 before the intervention to 17.98 (SD 1.25) after the intervention ($p = 0.001$), with statistically significant between-group differences reported for post-intervention scores across educational methods. However, standardized effect sizes and confidence intervals were not reported and could not be calculated reliably from the available summary data [17].

Parozzi et al. assessed performance-related clinical communication competence using a 26-item SBAR structured handover classification test in an uncontrolled pre–post quasi-experimental design. The mean score increased from 13.0 (SD 3.74) before the intervention to 16.2 (SD 4.28) after the intervention, corresponding to a mean difference of +3.2 points, a Cohen's *d* of 0.78, and a rank-biserial correlation of 0.414 ($p < 0.001$). Because this study did not include a control group and did not use individual-level pre–post matching, its findings were interpreted descriptively and not as comparative evidence of effectiveness [22].

3.14. Clinical Reasoning

Clinical reasoning outcomes were reported in four studies. Middeke et al. assessed clinical reasoning using a formative key-feature examination and a final EMERGE session. In the key-feature examination, the EMERGE group achieved a higher total score than the PBL group, with median scores of 62.5% (IQR 17.7) versus 54.2% (IQR 21.9), respectively. This corresponded to a median difference of +8.3 percentage points ($p = 0.015$). Standardized effect sizes and confidence intervals were not reported and could not be calculated reliably because results were presented as medians and interquartile ranges [20].

In a later randomized study, Middeke et al. assessed the transfer of clinical reasoning trained with EMERGE to comparable virtual patient cases. No significant difference was observed between the three randomized groups in the overall sum score across four cases. In exposed versus unexposed analyses, students previously exposed to related cases scored higher for NSTEMI cases ($67.7\% \pm 11.3$ vs. $58.8\% \pm 13.8$; $p = 0.019$; approximate SMD = 0.74) and asthma exacerbation cases ($65.0\% \pm 15.9$ vs. $58.8\% \pm 14.0$; $p = 0.015$; approximate SMD = 0.41). No significant differences were observed for pancreatitis or gastrointestinal hemorrhage [21].

Wu et al. evaluated the Virtual Dental Clinic as a serious game designed to support clinical reasoning, treatment planning, and instrument selection among clerkship dental students. In the quasi-experimental ex post facto component, students who used the VDC achieved higher scores on the national qualification test than those who did not use it. Mean qualification test scores were 70.00 (SD 13.54) among non-users, 80.00 (SD 5.59) among students who used VDC once, and 79.58 (SD 6.90) among those who used it at least twice ($p = 0.029$). When students with any VDC use were combined and compared with non-users, the mean difference was approximately +9.76 points, with an approximate Cohen's *d* of 1.01. VDC use was also associated with higher performance scores in periodontics ($p = 0.037$) and conservative dentistry ($p = 0.040$). However, because VDC use was self-selected and the study used a quasi-experimental ex post facto design, these findings

were interpreted cautiously as association-based evidence rather than causal evidence of effectiveness [24].

Ghafouri et al. also reported clinical reasoning-related outcomes using key-feature questions designed to assess students' ability to apply health assessment knowledge in clinical scenarios. Key-feature question scores improved from a mean of 12.90 before the intervention to 17.98 (SD 1.25) after the intervention ($p = 0.001$). However, standardized effect sizes and confidence intervals were not reported and could not be calculated reliably from the available summary data [17].

3.15. Secondary Outcomes

Table 5 summarizes the quantitative findings for secondary outcomes.

Knowledge Acquisition

In Tan et al. [23], knowledge scores differed significantly between groups ($p < 0.001$). In Mansoori et al. [19], knowledge scores were significantly higher in the VR group compared with the lecture group ($p = 0.019$).

In Yang and Oh [25], knowledge differed significantly across the three groups ($n = 29$, 28, and 26; $p < 0.05$). In Ghafouri et al. [17], knowledge outcomes showed significant between-group differences ($p < 0.05$). In Koivisto et al. [18], greater knowledge change was observed in the intervention group ($n = 140$) compared with the control group ($n = 136$), with $p < 0.05$.

3.16. Self-Efficacy

In Adhikari et al. (2021), self-efficacy improved significantly following the intervention ($n = 19$; $p < 0.001$) [13]. In Ghafouri (2024), self-efficacy differed significantly between study arms ($p < 0.05$) [17].

3.17. Motivation and Self-Confidence

In Yang (2022), motivation and self-confidence scores differed significantly between groups ($p < 0.05$) [25].

3.18. Satisfaction and Usability

In Alyami (2019) [14], satisfaction scores were significantly higher in the serious game group ($p < 0.001$). In Birrenbach (2021) [15] and Wu (2021) [24], usability and satisfaction measures were reported descriptively, with quantitative scale scores provided in the original articles. Parozzi (2025) [22] also assessed learner experience using the 33-item Player Experience Inventory (PXI), indicating that user experience and acceptability were evaluated as part of the intervention, although no detailed numerical PXI results were reported in the extracted table.

3.19. Anxiety

In Adhikari et al. (2021), anxiety scores decreased significantly after the intervention ($p < 0.001$) [13].

Table 5. Quantitative Results for Secondary Outcomes.

Study	Outcome Category	Assessment Tool	<i>n</i> (Intervention/Control)	Intervention (Mean $\pm$ SD or Median [IQR])	Control (Mean $\pm$ SD or Median [IQR])	Effect Estimate	95% CI	<i>p</i> -Value
[23]	Knowledge	Knowledge test	57/46	NR	NR	Significant difference reported	NR	<0.001
[23]	Confidence	Confidence scale	57/46	NR	NR	Significant difference reported	NR	<0.001
[14]	Satisfaction	Satisfaction questionnaire	27/19	NR	NR	Higher satisfaction in intervention	NR	<0.001
[19]	Knowledge	Learning outcome test	25/25	NR	NR	Significant difference reported	NR	0.019
[24]	Satisfaction	Usability/feedback scale	34 (users)	NR	NR	NR	NR	NR
[13]	Self-efficacy	NASC-CDM scale	19 (pre–post)	NR (post higher than pre)	NA	Pre–post difference	NR	<0.001
[13]	Anxiety	NASC-CDM scale	19 (pre–post)	NR (post lower than pre)	NA	Pre–post difference	NR	<0.001
[25]	Knowledge	Knowledge test	29/28/26	NR	NR	Significant group difference	NR	<0.05
[25]	Motivation	Motivation scale	29/28/26	NR	NR	Significant group difference	NR	<0.05
[25]	Self-confidence	Self-confidence scale	29/28/26	NR	NR	Significant group difference	NR	<0.05
[17]	Knowledge	Knowledge test	166 total	NR	NR	Significant difference reported	NR	<0.05
[17]	Self-efficacy	Self-efficacy scale	166 total	NR	NR	Significant difference reported	NR	<0.05
[18]	Knowledge	SNK test	140/136	NR	NR	Greater change in intervention group	NR	<0.05
[22]	SBAR handover performance (knowledge/skill)	Player Experience Inventory (PXI)	41/N/A	Positive mean scores across domains (e.g., Meaning $\approx$ 1.29–2.12, Enjoyment $\approx$ 1.51–1.56)	N/A	NR	NR	NR

3.20. Discussion

This systematic review included 13 studies evaluating serious games and virtual reality-based interventions in health sciences education [13–25]. Overall, the findings suggest that these interventions may support knowledge acquisition, learner confidence, satisfaction, motivation, and selected aspects of clinical reasoning. However, evidence for objective performance-based clinical competence was less consistent. This variability appears to reflect differences in intervention design, clinical domains, assessment tools, comparator conditions, and follow-up duration across studies.

Although several studies reported statistically significant differences favoring serious game- or VR-based interventions, statistical significance alone does not necessarily indicate educational or clinical importance. Therefore, positive findings were interpreted not only according to *p*-values, but also in relation to the magnitude of the observed differences, the outcome scale, the availability of standardized effect sizes or confidence intervals, follow-up results, and the methodological robustness of the studies. For example, some statistically significant findings represented modest absolute differences, such as a two-point improvement on a 17-item procedural checklist, whereas other findings showed larger standardized effects but were derived from uncontrolled or self-selected designs. Consequently, the educational significance of these findings should be interpreted cautiously, particularly when effect sizes, confidence intervals, long-term retention data, or direct links to real clinical performance were unavailable.

Positive findings should also be interpreted in light of the methodological limitations of the included evidence. Several studies had small sample sizes, heterogeneous intervention formats, variable comparator conditions, short follow-up periods, and moderate risk of bias. In addition, some findings were derived from non-randomized, feasibility, uncontrolled pre-post, or self-selected exposure designs, which limits causal inference. Therefore, although the overall direction of evidence suggests potential educational benefits of serious games and VR-based interventions, the strength of conclusions regarding effectiveness remains limited, particularly for objective performance-based clinical competence and long-term transfer to real clinical practice.

A key methodological feature of this review was the separation of lower-level cognitive outcomes (knowledge) from higher-level behavioral outcomes (performance-based competence), consistent with Miller's pyramid. In Miller's model, written tests typically reflect "knows," whereas OSCEs, structured simulations, and direct performance assessments more closely align with "shows how." This distinction is widely recognized in health professions education and supports the analytical decision to avoid treating knowledge-only outcomes as direct proxies for clinical performance [12].

Our findings are consistent with earlier systematic reviews reporting substantial heterogeneity in serious game interventions, populations, comparators, and outcome measures, which complicates pooled quantitative synthesis and often necessitates narrative synthesis approaches. For example, Gentry et al. reported heterogeneity across serious gaming/gamification studies in health professions education and noted variable effects across outcomes such as knowledge, skills, and satisfaction [6].

Maheu-Cadotte et al. emphasized that the effectiveness of serious games may vary by design features and outcome type, and highlighted methodological variability across included studies [26].

3.21. Outcome Measurement and Heterogeneity

A notable feature across the included studies is the diversity of primary outcome operationalization. Performance-based competence was measured through different formats (OSCE scoring, checklist-based simulation performance, time-to-performance thresholds

and structured handover performance), while clinical reasoning was assessed using key-feature examinations, transfer tasks, and diagnostic reasoning-related assessments. This measurement heterogeneity limits direct comparability even within the same outcome category and likely contributes to variability in reported effects across studies. Consistent reporting of measurement validity (e.g., evidence for instrument validity, rater training, inter-rater reliability) remains critical for interpreting results within and across studies and is a recurring methodological consideration in the serious games evidence base, as highlighted in prior reviews [26].

3.22. Educational and Curricular Implications

The included studies show that serious games and virtual reality-based interventions have been applied across several areas of health sciences education, including resuscitation, diagnostic reasoning, procedural skills training, structured handover, and nursing decision-making. These interventions varied in format, ranging from non-digital board games to desktop-based serious games and immersive virtual reality systems. Rather than representing a single instructional modality, serious games should therefore be understood as a family of scenario-based, interactive learning strategies that can be adapted to different competency targets and educational contexts.

From a curricular design perspective, effective integration requires constructive alignment between intended learning outcomes, instructional activities, and assessment methods. In competency-based education frameworks, alignment is essential to ensure that teaching strategies support measurable competency development [26]. This reinforces the need to align serious game design with the intended competency level and the corresponding assessment method.

Importantly, the educational value of serious games appears to depend less on technological immersion alone than on instructional design quality. Structured clinical scenarios, active decision-making, and embedded feedback mechanisms were common characteristics across interventions targeting higher-order competencies. Prior syntheses emphasize that these pedagogical components rather than technological sophistication alone are central to serious game effectiveness in health sciences education [6,26].

This suggests that curricular planners should prioritize scenario authenticity, feedback design, and competency alignment when integrating serious games into training programs.

Finally, the variability in exposure duration and integration models across studies highlights different implementation strategies, ranging from adjunctive preparatory tools to embedded multi-session modules. For institutions adopting competency-based medical or nursing education models, serious games may function as complementary platforms for deliberate practice, structured rehearsal, or pre-simulation preparation. However, their curricular positioning should be explicitly defined in relation to targeted competency milestones and validated assessment frameworks to ensure coherence within programmatic assessment structures.

Several limitations emerge from the included studies and reporting patterns. First, outcome measures were not standardized, and confidence intervals or standardized effect sizes were not consistently reported, limiting cross-study quantitative comparability. Second, follow-up assessments beyond immediate post-intervention were limited, restricting inferences about retention and transfer over time. In addition, publication bias, small-study effects, and technological novelty bias should be considered when interpreting the findings. Because educational technology studies with positive or innovative results may be more likely to be published, the available evidence may overrepresent favorable findings. Several included studies had small sample sizes or pilot designs, which may increase the risk of imprecise or inflated effect estimates. Moreover, learner satisfaction, motivation,

and engagement may partly reflect the novelty or attractiveness of serious games and VR-based tools rather than sustained educational effectiveness. This is particularly relevant when outcomes are measured immediately after the intervention and when long-term retention, transfer to clinical practice, or objective performance-based outcomes are not assessed. Third, several studies emphasized secondary outcomes, such as satisfaction, usability, motivation, or knowledge-only measures, which are informative but should be interpreted within the lower tiers of competency frameworks. Finally, non-randomized and feasibility designs, while valuable for early-stage development and acceptability, contribute primarily to descriptive synthesis rather than comparative effectiveness estimates.

3.23. Recommendations for Future Research

Future trials would benefit from (1) clearer alignment between competency targets and validated assessment strategies (particularly for performance-based competence and clinical reasoning), (2) consistent reporting of quantitative results (means/SDs or medians/IQRs, effect estimates, and confidence intervals), (3) inclusion of retention and transfer outcomes with longer follow-up, and (4) stronger reporting transparency for risk-of-bias domains (e.g., allocation processes, assessor blinding, missing data handling). Additionally, consensus on a core outcome set for serious games in health sciences education would improve comparability and strengthen evidence synthesis.

3.24. Conclusions

This systematic review suggests that serious games and virtual reality-based interventions may represent promising complementary educational strategies in health sciences education. The most consistent findings were observed for knowledge acquisition, learner engagement, self-efficacy, satisfaction, motivation, and selected clinical reasoning outcomes. However, evidence for objective improvement in performance-based clinical competence remains variable and should be interpreted cautiously because of heterogeneity in intervention design, clinical domains, comparator groups, assessment methods, follow-up duration, sample size, and risk of bias across the included studies.

These findings suggest that serious games may be most useful when integrated into competency-based curricula as structured adjuncts to conventional teaching, simulation, or clinical training, rather than as standalone replacements. Their potential educational value appears to depend not only on technological features, but also on scenario authenticity, active learner decision-making, feedback design, and alignment with validated competency-based assessment frameworks.

Future studies should use more rigorous designs, standardized outcome frameworks, validated performance-based assessments, transparent reporting of effect estimates, standardized effect sizes and confidence intervals, and longer follow-up periods to evaluate retention and transfer to simulated or real clinical practice. Larger randomized controlled trials are needed before stronger conclusions can be drawn regarding the effectiveness of serious games and VR-based interventions for objective performance-based clinical competence.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ime5020055/s1>, Supplementary Material S1: Full Electronic Search Strategies.

Author Contributions: K.A. contributed to conceptualization, methodology, formal analysis, investigation, and writing of the original draft. M.A.B. contributed to conceptualization, methodology, investigation, formal analysis, and manuscript revision. H.N. acted as the third reviewer, supervised the study, and critically reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analyzed during this systematic review are included in this article and its Supplementary Materials.

Acknowledgments: The authors used an artificial intelligence tool for language editing and text refinement only. The AI tool was not involved in study selection, data extraction, analysis, or interpretation. The authors take full responsibility for the integrity and accuracy of the work.

Conflicts of Interest: The authors declare no conflicts of interests.

References

1. World Health Organization. *Transforming and Scaling Up Health Professionals' Education and Training: World Health Organization Guidelines 2013*; World Health Organization: Geneva, Switzerland, 2013; p. 124.
2. Hui, T.; Zakeri, M.A.; Soltanmoradi, Y.; Rahimi, N.; Hossini Rafsanjanipoor, S.M.; Nouroozi, M.; Dehghan, M. Nurses' Clinical Competency and Its Correlates: Before and during the COVID-19 Outbreak. *BMC Nurs.* **2023**, *22*, 156. [[CrossRef](#)]
3. El Amrani, M.; Baba, M.A.; Nassik, H. Impact of Simulation-Based Education on the Development of Non-Technical Skills in Health Sciences Students: A Systematic Review. *Int. Med. Educ.* **2026**, *5*, 45. [[CrossRef](#)]
4. Frenk, J.; Chen, L.; Bhutta, Z.A.; Cohen, J.; Crisp, N.; Evans, T.; Fineberg, H.; Garcia, P.; Ke, Y.; Kelley, P. Health Professionals for a New Century: Transforming Education to Strengthen Health Systems in an Interdependent World. *Lancet* **2010**, *376*, 1923–1958. [[CrossRef](#)] [[PubMed](#)]
5. Boyle, E.A.; Hainey, T.; Connolly, T.M.; Gray, G.; Earp, J.; Ott, M.; Lim, T.; Ninaus, M.; Ribeiro, C.; Pereira, J. An Update to the Systematic Literature Review of Empirical Evidence of the Impacts and Outcomes of Computer Games and Serious Games. *Comput. Educ.* **2016**, *94*, 178–192. [[CrossRef](#)]
6. Gentry, S.V.; Gauthier, A.; Ehrstrom, B.L.; Wortley, D.; Lilienthal, A.; Car, L.T.; Dauwels-Okutsu, S.; Nikolaou, C.K.; Zary, N.; Campbell, J. Serious Gaming and Gamification Education in Health Professions: Systematic Review. *J. Med. Internet Res.* **2019**, *21*, e12994. [[CrossRef](#)] [[PubMed](#)]
7. Demircan, B.; Kiyak, Y.; Kaya, H. The Effectiveness of Serious Games in Nursing Education: A Meta-Analysis of Randomized Controlled Studies. *Nurse Educ. Today* **2024**, *142*, 106330. [[CrossRef](#)] [[PubMed](#)]
8. Min, A.; Min, H.; Kim, S. Effectiveness of Serious Games in Nurse Education: A Systematic Review. *Nurse Educ. Today* **2022**, *108*, 105178. [[CrossRef](#)] [[PubMed](#)]
9. Lee, M.; Shin, S.; Lee, M.; Hong, E. Educational Outcomes of Digital Serious Games in Nursing Education: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. *BMC Med. Educ.* **2024**, *24*, 1458. [[CrossRef](#)] [[PubMed](#)]
10. Dickson, K.; Yeung, C.A. PRISMA 2020 Updated Guideline. *Br. Dent. J.* **2022**, *232*, 760–761. [[CrossRef](#)] [[PubMed](#)]
11. Witheridge, A.; Ferns, G.; Scott-Smith, W. Revisiting Miller's Pyramid in Medical Education: The Gap between Traditional Assessment and Diagnostic Reasoning. *Int. J. Med. Educ.* **2019**, *10*, 191–192. [[CrossRef](#)] [[PubMed](#)]
12. Adhikari, R.; Kydonaki, C.; Lawrie, J.; O'Reilly, M.; Ballantyne, B.; Whitehorn, J.; Paterson, R. A Mixed-Methods Feasibility Study to Assess the Acceptability and Applicability of Immersive Virtual Reality Sepsis Game as an Adjunct to Nursing Education. *Nurse Educ. Today* **2021**, *103*, 104944. [[CrossRef](#)] [[PubMed](#)]
13. Alyami, H.; Alawami, M.; Lyndon, M.; Alyami, M.; Coomarasamy, C.; Henning, M.; Hill, A.; Sundram, F. Impact of Using a 3D Visual Metaphor Serious Game to Teach History-Taking Content to Medical Students: Longitudinal Mixed Methods Pilot Study. *JMIR Serious Games* **2019**, *7*, e13748. [[CrossRef](#)] [[PubMed](#)]
14. Birrenbach, T.; Zbinden, J.; Papagiannakis, G.; Exadaktylos, A.K.; Müller, M.; Hautz, W.E.; Sauter, T.C. Effectiveness and Utility of Virtual Reality Simulation as an Educational Tool for Safe Performance of COVID-19 Diagnostics: Prospective, Randomized Pilot Trial. *JMIR Serious Games* **2021**, *9*, e29586. [[CrossRef](#)] [[PubMed](#)]
15. Drummond, D.; Delval, P.; Abdenouri, S.; Truchot, J.; Ceccaldi, P.-F.; Plaisance, P.; Hadchouel, A.; Tesnière, A. Serious Game versus Online Course for Pretraining Medical Students before a Simulation-Based Mastery Learning Course on Cardiopulmonary Resuscitation: A Randomised Controlled Study. *Eur. J. Anaesthesiol.* **2017**, *34*, 836–844. [[CrossRef](#)] [[PubMed](#)]
16. Ghafouri, R.; Zamanzadeh, V.; Nasiri, M. Comparison of Education Using the Flipped Class, Gamification and Gamification in the Flipped Learning Environment on the Performance of Nursing Students in a Client Health Assessment: A Randomized Clinical Trial. *BMC Med. Educ.* **2024**, *24*, 949. [[CrossRef](#)] [[PubMed](#)]
17. Koivisto, J.-M.; Buure, T.; Engblom, J.; Rosqvist, K.; Haavisto, E. The Effectiveness of Simulation Game on Nursing Students' Surgical Nursing Knowledge—A Quasi-Experimental Study. *Teach. Learn. Nurs.* **2024**, *19*, e22–e29. [[CrossRef](#)]

18. Mansoori, M.S.; Khazaei, M.R.; Azizi, S.M.; Niromand, E. Comparison of the Effectiveness of Lecture Instruction and Virtual Reality-Based Serious Gaming Instruction on the Medical Students' Learning Outcome about Approach to Coma. *BMC Med. Educ.* **2021**, *21*, 347. [[CrossRef](#)] [[PubMed](#)]
19. Middeke, A.; Anders, S.; Schuelper, M.; Raupach, T.; Schuelper, N. Training of Clinical Reasoning with a Serious Game versus Small-Group Problem-Based Learning: A Prospective Study. *PLoS ONE* **2018**, *13*, e0203851. [[CrossRef](#)] [[PubMed](#)]
20. Middeke, A.; Anders, S.; Raupach, T.; Schuelper, N. Transfer of Clinical Reasoning Trained with a Serious Game to Comparable Clinical Problems: A Prospective Randomized Study. *Simul. Healthc.* **2020**, *15*, 75–81. [[CrossRef](#)] [[PubMed](#)]
21. Parozzi, M.; Meraviglia, I.; Ferrara, P.; Morales Palomares, S.; Mancin, S.; Sguanci, M.; Lopane, D.; Destrebecq, A.; Lusignani, M.; Mezzalana, E.; et al. Effectiveness of a Gamification-Based Intervention for Learning a Structured Handover System Among Undergraduate Nursing Students: A Quasi-Experimental Study. *Nurs. Rep.* **2025**, *15*, 322. [[CrossRef](#)] [[PubMed](#)]
22. Tan, A.J.Q.; Lee, C.C.S.; Lin, P.Y.; Cooper, S.; Lau, L.S.T.; Chua, W.L.; Liaw, S.Y. Designing and Evaluating the Effectiveness of a Serious Game for Safe Administration of Blood Transfusion: A Randomized Controlled Trial. *Nurse Educ. Today* **2017**, *55*, 38–44. [[CrossRef](#)] [[PubMed](#)]
23. Wu, J.-H.; Du, J.-K.; Lee, C.-Y. Development and Questionnaire-Based Evaluation of Virtual Dental Clinic: A Serious Game for Training Dental Students. *Med. Educ. Online* **2021**, *26*, 1983927. [[CrossRef](#)] [[PubMed](#)]
24. Yang, S.-Y.; Oh, Y.-H. The Effects of Neonatal Resuscitation Gamification Program Using Immersive Virtual Reality: A Quasi-Experimental Study. *Nurse Educ. Today* **2022**, *117*, 105464. [[CrossRef](#)] [[PubMed](#)]
25. Maheu-Cadotte, M.-A.; Cossette, S.; Dubé, V.; Fontaine, G.; Mailhot, T.; Lavoie, P.; Cournoyer, A.; Balli, F.; Mathieu-Dupuis, G. Effectiveness of Serious Games and Impact of Design Elements on Engagement and Educational Outcomes in Healthcare Professionals and Students: A Systematic Review and Meta-Analysis Protocol. *BMJ Open* **2018**, *8*, e019871. [[CrossRef](#)] [[PubMed](#)]
26. Biggs, J. Enhancing Teaching through Constructive Alignment. *High Educ.* **1996**, *32*, 347–364. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.