SkillsVR™

STUDY

Integrating AI with VR for Post-Operative Assessment Training in Healthcare

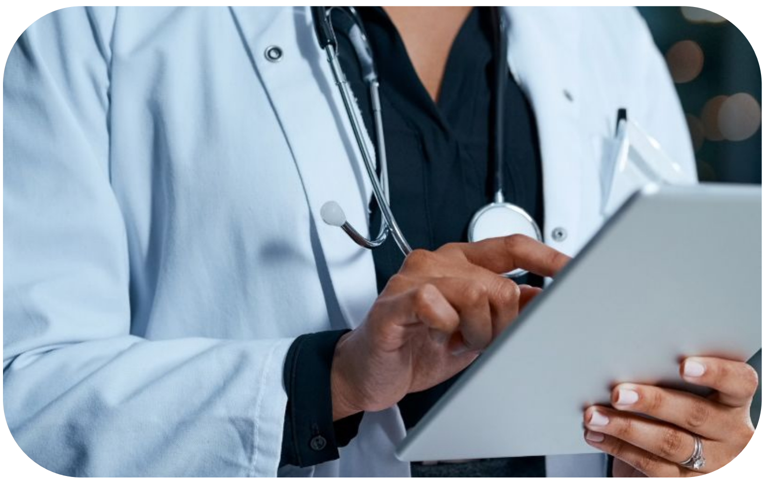
www.skillsvr.com

Introduction

The integration of immersive technologies into medical education has gained increasing traction in recent years, with virtual reality (VR) providing learners with a safe, scalable, and repeatable environment to practice clinical skills (Cook et al., 2012). VR-based simulation has been particularly effective in domains such as surgical training, procedural practice, and patient communication, offering experiential learning without exposing patients to risk (Seymour et al., 2002; Kyaw et al., 2019). Despite these advances, many VR applications in healthcare education remain limited to replicating physical interactions and procedural steps, without systematically capturing or analyzing learner reasoning and communication behaviors.

In parallel, artificial intelligence (AI) has demonstrated growing potential in education and healthcare, particularly in natural language processing, automated assessment, and decision-support systems (Holmes et al., 2019; Zawacki-Richter et al., 2019). AI has been leveraged to provide personalized feedback, detect learner errors, and identify behavioral patterns that might otherwise go unnoticed (Topol, 2019; Davenport & Kalakota, 2019). Yet the intersection of VR and AI in medical training is still at an early stage. While VR provides authentic experiential environments, AI offers the capacity to analyze learner performance either in real time or post hoc, producing structured insights that can inform reflection and coaching.

Current VR systems in medical education emphasize skill replication and procedural accuracy but rarely integrate automated performance analysis, bias detection, or structured reflective feedback. This paper addresses this gap by examining the integration of AI into a VR-based post-operative patient assessment module, developed in collaboration with SkillsVR and Northwell Health. Specifically, the study demonstrates how AI-driven speech recognition, clinical checklists, and automated reporting can extend VR training from simulation toward data-rich, reflection-first pedagogy, while also highlighting scalability challenges, potential for bias detection, and opportunities to support more consistent and informed clinical decision-making.



2. Background / Related Work

2.1 VR in Healthcare Education

VR has evolved from a novel technology to a credible modality for healthcare training. Early work in surgical education demonstrated that VR-trained residents performed procedures faster and with fewer errors compared to traditionally trained peers (Seymour et al., 2002). Meta-analyses of technology-enhanced simulation have reinforced its effectiveness, reporting moderate-to-large improvements in knowledge, skills, and behavior (Cook et al., 2012; Haerling, 2018). Beyond procedural training, VR has been successfully applied in areas such as teamwork, communication, and emergency management, leveraging immersion and embodiment to approximate real-world cognitive and affective demands (Kozan et al., 2019; Kyaw et al., 2019; Almarzooq et al., 2020). However, most VR implementations continue to focus on task replication, while structured analysis of learner reasoning and decision-making remains limited.

2.2 AI for Assessment and Feedback in Learning

AI, and in particular natural language processing (NLP), has become increasingly relevant in education. Studies have shown that AI can automate assessment, provide individualized feedback, and detect patterns in learner performance (Zawacki-Richter et al., 2019; Holmes et al., 2019). In healthcare contexts, AI has been used to evaluate diagnostic reasoning, score clinical notes, and align learner performance with competency frameworks (Davenport & Kalakota, 2019; Chan & Zary, 2019). Importantly, AI can transform unstructured learner input—such as text or speech—into structured datasets, supporting post-hoc reflection and faculty debriefing (Topol, 2019). Yet few systems have integrated this analysis directly into immersive simulations, where learner discourse and reasoning unfold naturally.

2.3 Bias in Clinical Decision-Making and Training

A large body of research highlights disparities in clinical assessment and treatment across race, gender, and language (Smedley et al., 2003; Green et al., 2007; Hall et al., 2015). Implicit bias has been shown to influence diagnostic urgency, pain assessment, and therapeutic decisions, often independent of explicit attitudes (Sabin et al., 2009; Chapman et al., 2013). Educational interventions targeting bias frequently rely on workshops or DEI-oriented modules, which risk performative responses and limited long-term impact (Noon, 2018). Simulation has been proposed as an alternative, enabling observation of clinical behavior under realistic conditions. However, when scenarios explicitly highlight "bias," learners may adapt their responses. Embedding diversity naturally into routine assessments, followed by AI-driven post-session analysis, may provide a more authentic pathway to bias awareness.

2.4 Integrating VR and AI for Data-Rich, Authentic Learning

Recent studies have explored combining VR simulations with AI analytics to generate actionable feedback. Prototype systems have used multimodal data (speech, actions, physiological signals) to provide dashboards or narrative feedback to coaches (Chen et al., 2020; Liu et al., 2022; Luxton, 2016). However, much of this work emphasizes real-time prompts or performance support, which can alter authentic learner behavior. Few systems have adopted a reflection-first approach that defers analysis until after the simulation, while also aligning outputs with clinical frameworks (e.g., Aldrete scoring). This motivates the reflection-first approach evaluated in this study.

3. Methodology

3.1 Project Context

The project was conducted through a collaboration between SkillsVR and Northwell Health, aimed at designing a VR-based training module that both replicates routine post-operative patient assessments and integrates AI to analyze learner interactions. The module was implemented in a hospital-inspired virtual environment, where learners engaged with three distinct patient scenarios. Each scenario was developed to reflect realistic clinical encounters while introducing demographic diversity and potential clinical complications. The patients included:

1. Brian Stewart (White, Male, 59) – Routine recovery with normal vitals.
2. Cheryl Robinson (Black, Female, 63) – Post-surgical hypotension requiring recognition and clinical judgment.
3. Shruthi Venkatasubramanian (Asian, Female, 61) – Cognitive impairment and communication challenges with otherwise stable vitals.

These scenarios provided both a baseline (normal recovery) and cases with subtle or overt abnormalities, allowing for comparison of learner decision-making across contexts. Importantly, demographic variables such as race, gender, and communication ability were embedded into the cases to enable analysis of unconscious bias in clinical assessment.

3.2 Technical Workflow

The technical pipeline of the module consisted of four main stages:

1. **Speech-to-Text Capture**
   Learners interacted with patients verbally throughout the simulation. All spoken input was captured and transcribed using speech-to-text models.
2. **AI-Driven Clinical Analysis**
   The transcribed data was mapped against structured clinical checklists derived from post-operative assessment protocols, including vital sign interpretation, pain assessment, cognitive evaluation, and the Aldrete scoring system.

3. Bias Analysis

   Learner responses were analyzed across the three patient profiles to identify potential disparities in decision-making. Metrics included time spent per task, diagnostic choices, and differences in urgency or follow-up actions. Subtle communication indicators (e.g., addressing fatigue, confusion, or patient discomfort) were also examined for consistency.

4. Automated Report Generation

   At session end, the AI generated structured reports (PDF, CSV) summarizing performance, highlighting missed or incomplete steps, and providing data points for further analysis. Reports were designed for post-session debriefs, supporting reflective learning rather than immediate corrective feedback. While the AI organized learner interactions, it did not apply bias baselines. Instead, Dr. Robert and the Northwell team reviewed outputs to analyze patterns suggesting bias (e.g., differences in urgency, completeness, or diagnostic assumptions across demographics). This ensured that bias insights emerged from expert interpretation of structured data rather than pre-programmed benchmarks.

3.3 Design Principles

The module was explicitly designed to prioritize authentic learner behavior. Three design principles guided its development:

- No real-time feedback: Ensuring learners' actions reflected genuine decision-making without adaptation to perceived monitoring.
- Reflection-first approach: Positioning AI reports as tools for structured post-session coaching.
- Authentic scenario flow: Embedding assessments in routine clinical contexts so that bias-related behaviors emerge naturally.

3.4 Data Sources

The primary data sources included:

- **Learner speech** (captured and transcribed).
- **Clinical decision points** (e.g., classification of vitals, Aldrete scoring, discharge recommendations).
- **Performance metadata** (time taken, diagnostic pathways, communication style).

This enabled quantitative analysis (checklists, timing) and qualitative insights (communication patterns, interpretation of ambiguous symptoms).

Northwell clinicians developed baseline patient profiles (vitals, recovery status, communication challenges) serving as the reference standard for comparing learner assessments with expected "ideal" assessments. For three structured questions per learner, clinicians also created ideal model responses. The AI compared learner responses against these model answers to assess completeness, accuracy, and clinical reasoning, anchoring automated reports in expert-designed criteria.

4. Results / Findings

4.1 Accuracy of AI Capture

Across all scenarios, speech-to-text reliably transcribed learner responses, with minimal errors in clinical terminology. AI analysis consistently identified whether learners addressed key clinical checks (blood pressure, heart rate, oxygen saturation, respiration, pain assessment, cognitive evaluation). For example, in Brian Stewart's case (normal recovery), learners who verbalized "BP 117/63, HR 84, O_2 sat 98%" received full credit for vitals recognition, whereas omission of respiratory rate was flagged.

The Aldrete scoring process was accurately processed. When learners identified a total score of 10/10 for Brian Stewart, reports validated the scoring and confirmed readiness for Phase II recovery. Miscalculations or weak rationale were flagged for coaching follow-up.

4.2 Value of AI-Generated Reports

One of the central outcomes was the utility of AI-generated reports as reflective learning tools. Reports provided both quantitative scoring and qualitative insights into communication style and decision-making.

- Quantitative scoring: Automated checklists validated whether learners performed each required step (e.g., assessing pain, asking orientation questions, interpreting vitals).
- Qualitative feedback: Reports highlighted learner tendencies, such as hesitancy in responding to abnormal findings, or omission of cognitive assessment in scenarios where mental status was critical.

In Cheryl Robinson's case (hypotension), learners who downplayed circulatory compromise were identified via comments such as "Circulatory instability identified but not explicitly linked to risk of deterioration," providing focused prompts for debrief.

4.3 Bias Detection Insights

Because scenarios were not framed as DEI exercises, learners interact authentically, allowing unconscious biases to surface.

- Case 1 – Brian Stewart (White male): Learners tended to provide confident, complete assessments, often scoring him quickly as stable for discharge.
- Case 2 – Cheryl Robinson (Black female): Despite overt hypotension, some learners hesitated to escalate or underestimated the clinical significance of her abnormal vitals. This suggested a lower perceived urgency, a finding consistent with documented disparities in the treatment of Black patients [REF].
- Case 3 – Shruthi Venkatasubramanian (Asian female, cognitive impairment): Learners frequently attributed her repetitive speech to a "language barrier" rather than considering neurological complications such as stroke or delirium. This highlighted a potential diagnostic bias in interpreting communication differences as cultural rather than medical.

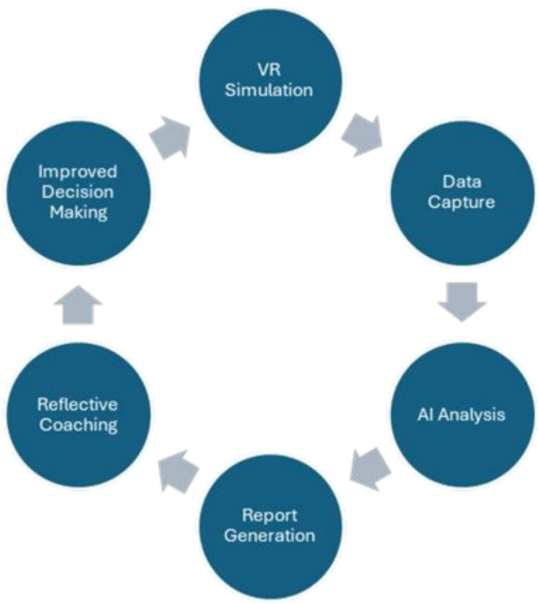
By comparing learner responses across cases, the AI system was able to flag differences in urgency, diagnostic reasoning, and assumptions, producing bias indicators that could be explored in post-training reflection.

4.4 Scalability and Cost Implications

AI tokenization created both opportunities and challenges. The system is scaled by generating structured reports with minimal instructor input, but recurring token costs rose with learner volume and session length. For large cohorts, bulk transcription and analysis would materially increase usage, underscoring the need to balance educational value with financial sustainability across medical schools, residency programs, and hospital training units.

AI-Augmented Immersive Learning Framework (Conceptual Model)

- **Stage 1: Immersive Clinical Simulation (VR Environment).** Authentic, diverse, clinically varied scenarios; focus on clinical decision-making, not DEI "performance."
- **Stage 2: Data Capture (Within Simulation).** Speech-to-text; performance metadata (timing, pathways, engagement); clinical checklists (vitals, pain, cognition, Aldrete).
- **Stage 3: AI Analysis.** Checklist validation; bias indicators across demographics; alignment with expected outcomes.
- **Stage 4: Automated Report Generation.** PDF/CSV outputs covering checklist performance, decision rationale, communication features, potential bias patterns.
- **Stage 5: Reflective Learning & Coaching.** 1:1 or group debrief; reflection on decision accuracy, bias awareness, reasoning process; emphasis on self-awareness and continuous improvement.
- **Stage 6: Outcomes & Contributions- Improved Decision Making.** Educational impact (accuracy, structure); equity impact (bias awareness); scalability (applicable across healthcare domains); challenges (token costs, real-time feedback trade-offs, SME oversight).



5. Discussion

This study shows how integrating AI into VR-based medical training extends immersive simulation into a data-rich learning ecosystem. Naturalistic patient interactions paired with post-session AI analysis allowed authentic engagement during assessments and produced structured outputs for reflection and coaching.

5.1 Theoretical and Pedagogical Implications

The framework positions VR not only as a medium for experiential practice but also as a data-capture environment where decisions can be systematically analyzed. The reflection-first design, eschewing real-time corrective prompts, preserved authentic reasoning. Post-session reports then grounded structured coaching, aligning with research on delayed reflection for metacognition and bias recognition [REF]. AI thus augments, rather than replaces, human debriefing.

5.2 Bias Awareness Through Naturalistic Scenarios

Embedding diversity within routine assessments avoided priming performative behavior. Patterns observed across scenarios indicated differential urgency and attribution that, when reviewed post-session, supported bias-aware reflection while maintaining patient-centered focus.

5.3 Practical Considerations and Challenges

First, token-based AI services create recurring costs that scale with learner volume and scenario complexity; sustainable funding models are needed for large or distributed programs.

Second, although speech-to-text was reliable overall, accent and language variation remain limitations; model adaptation may be needed to ensure equitable accuracy.

Finally, AI outputs should be treated as decision-support, not definitive evaluations; educator judgment is essential to contextualize findings and avoid over-reliance on automated scoring.

5.4 Future Directions

Selective real-time feedback within VR warrants exploration, provided authenticity of decision-making is preserved. Future work should examine longitudinal impacts (e.g., transfer to clinical practice, reduction of bias-related disparities) and conduct cross-institutional studies to assess generalizability across diverse systems and learner populations.

6. Contributions

This paper makes several contributions to research and practice in medical education and immersive technologies:

- **Framework.** An AI-augmented immersive learning framework combining VR simulation with AI-driven speech analysis, checklist validation, and automated reporting.
- **Pedagogy.** A reflection-first design that preserves authenticity and enables structured debriefing.
- **Equity.** A method for surfacing unconscious bias via naturally embedded demographic diversity and post-session analysis.
- **Scaling guidance.** Practical considerations for deployment (token costs; speech-to-text accuracy).
- **Transferability.** Applicability beyond medicine to other communication- and decision-intensive domains.

7. Limitations

First, the AI provided post-session analysis only; future work may test limited real-time feedback without undermining authenticity.

Second, speech-to-text accuracy varied by accent and background noise; adaptation may be needed to ensure fairness.

Third, bias indicators require careful interpretation; educators remain central for contextualization.

Finally, the study used three scenarios and lacked longitudinal tracking; future studies should evaluate sustained effects on clinical reasoning and bias reduction.

8. Conclusion

Integrating AI with VR transformed simulation into a data-rich, reflection-oriented ecosystem that captures clinical reasoning, surfaces potential bias patterns, and generates coachable outputs. Remaining challenges—token costs, transcription equity, and the need for educator judgment—define practical next steps. The framework is scalable and adaptable beyond healthcare, offering a path toward more consistent, informed clinical decision-making and equity-aware training.

SkillsVR™

JBA Group Inc. US
3101 East 7th, PKWY
Denver, CO 80206
+1(650) 730–8836

Auckland, NZ Office
Level 3/24 Balfour Road, Parnell
Auckland, New Zealand 1052
0800 020 005

contact@skillsvr.com