

Original Research

Received 02 June 2026

Revised 10 Aug 2026

Accepted 01 Oct 2026

Natural Resources for Human Health 2026, Vol. 6 (13s): 1554–1567

KEYWORDS:

Medical Labeling, Instructions for Use, AI-Assisted Authoring, Virtual Reality, Human-Computer Interaction, Change-Impact Analysis, Regulatory Compliance, Immersive Guidance

From Labels to Immersive (VR) Guidance: Evaluating AI-Assisted Medical Labeling and Immersive IFUs through HCI

Himaja Challa¹, Avani Rao Alladi², Romit Soley³

¹User Experience Designer Edwards LifeSciences Seattle, WA

himaja.challa@gmail.com

²UX Designer Edwards Lifesciences Irvine, CA

alladiavanirao@gmail.com

³Product Designer Exxat Warren, NJ

soleyromit@gmail.com

ABSTRACT: Instructions for Use (IFUs) for transcatheter heart valve (THV) and transcatheter mitral and tricuspid therapy (TMFT) devices are among the most procedurally dense artifacts in medicine – each one must pass through design control, risk analysis, multilingual translation, regulatory submission and post-market surveillance. Two interventions for the same THV/TMFT labelled corpus were evaluated in this paper: (1) AI-assisted authoring and change-impact analysis for labeling content with applicable labeling governing standards, and (2) immersive, procedure-anchored IFU in virtual reality. Representative users assessed the accuracy, completeness, regulatory compliance, traceability, error detection, review efficiency, and usability of the performance of the AI approaches compared to the human approaches. They also completed IFU-related procedures on the device with both the traditional IFU and the immersive guidance experience. Follow up interviews were conducted to see the effect of immersion on comprehension, confidence, attention and perceived authority in labelling information. The study provides a framework for end-to-end HCI, linking the upstream issue of governance around label design to the downstream issue of procedural understanding. The framework does not encourage the use of AI or VR as alternative labels but rather as layers of augmentation that are traceable and under human supervision.

1. INTRODUCTION

Medical device labelling is an interesting field that sits at the crossroads of clinical practice, regulatory governance and human factors engineering. Instructions for Use (IFUs) are among the most complex procedural documents in the world of healthcare for transcatheter heart valve (THV) devices and transcatheter mitral and tricuspid therapy (TMFT) devices. Every IFU has to go through an iterative development process that calls for inputs from design control, risk analysis, multilingual translation, regulatory submission to various jurisdictions, and post-market surveillance requirements. The labeling ambiguity of such documents can lead to clinical errors during deployment, such as sizing, sequencing or not knowing when the device is contraindicated during procedural decision making.

The labelling regulation for medical devices has become more rigorous. Combined, the European Union Medical Device Regulation (EU MDR 2017/745), the U.S. FDA Unique Device Identification (UDI) system, and the growing harmonization efforts underway internationally under IMDRF have all increased the demands set on the accuracy, traceability, and change management of product labelling. In addition to being clinically accurate, manufacturers need to ensure that label content is internally consistent with the manufacturer's design and risk documentation, and is responsive to post-marketing intelligence, both as the label changes and as it is revised over time. The requirements pose a labeling governance challenge, which is essentially an information-management issue, and are ideal for computational augmentation.

The practical challenges of deploying AI systems within regulated enterprise workflows have received growing attention: Uddandarao (2026a) documents the gap between the promise of AI dashboards and their real-world implementation performance, a concern that directly applies to AI-assisted labeling systems operating within medical device governance workflows.

The potential of AI for regulatory document analysis is evident, with large language models (LLMs) and natural language processing (NLP) pipelines showing great promise. The development of AI systems has already been demonstrated to support compliance with various standards, like ISO 15223-1 for symbols, IEC 62366-1 for usability engineering, and FDA guidance on content requirements for labelling. Most current methods focus on a one-off document analysis labeling task, however, without taking into account the iterative and dynamic nature of labeling workflows, whereby a change to one section may trigger several dependent sections, languages, and regulatory submissions.

The labeling problem is only the first part of the puzzle; what happens to the IFU content as it is used in the procedure decision making process is the second part. Standard paper or PDF-based or standard PDF IFUs contain linear descriptions of procedures, forcing the clinician to mentally interpret the three-dimensional procedure from the two dimensional description. This mismatch between the information provided on a label and what is required in the actual procedure can lead to errors in use, especially when time is limited or when the procedure is being performed as part of training.

Virtual reality (VR) can serve as a possible link between labeling and procedural understanding. Finally, IFUs can be rendered in an immersive way that physically places IFU information in the context of spatial, interactive, and 3D representations of device anatomy and deployment sequences, which may help ease the cognitive burden of converting text instructions into procedural steps. Previous research on HCI has shown that spatial memory, embodied interaction and contextual cueing in immersive environments can facilitate procedural learning and recall. No previous study has related upstream AI-assisted labeling governance to the downstream immersive IFU delivery in the same framework, though.

In this paper, this is remedied by presenting a mixed methods evaluation of two interventions that were applied to the same THV/TMTT labeling corpus. The first intervention utilizes AI-assisted authoring and change-impact analysis to label content, and assesses performance based on the review conducted on its own against the content reviewed by humans, across seven criteria: accuracy, completeness, regulatory compliance, traceability, error-detection capability, review efficiency and usability. The second intervention is identical in its procedural content to an immersive VR IFU but its impact on procedural task performance, comprehension, confidence and perceived authority is compared with traditional forms of IFUs. Follow-up qualitative interviews explore how immersion affects perceived authority and trustworthiness of information labeled. The paper presents an end-to-end HCI framework, bridging from upstream, the control of labeling governance, to downstream, the procedural understanding of labelling, to illustrate how AI and VR can act as traceable, human-supervised augmentation layers rather than substitutes for approved labelling.

2. LITERATURE REVIEW

2.1 Medical Device Labeling Governance

There are numerous and overlapping Medicare device labeling requirements. The MDR 2017/745 sets out comprehensive IFU content, language and revision management requirements in the EU, and ISO 15223 defines symbols for use in labelling and ISO 62366-1 outlines a framework for usability engineering applied to labelling design. There are additional requirements in the United States based on FDA regulations as well as new guidance for electronic IFUs (eIFUs) under 21 CFR Part 801. In an effort to minimize duplication between jurisdictions, IMDRF has worked towards harmonizing the requirements, but manufacturers who operate on a global level still need to comply with the various jurisdictions for the content, format and changes of labels. This is especially the case with regard to change management. If a design change, risk update or post market corrective action requires a change to the IFU, this may ripple through several parts of the IFU, several translations, multiple regulatory submissions, and multiple downstream documents including training materials and quick-reference guides. Manual dependence tracking can be inefficient and prone to human mistakes, which can have a negative impact on patient safety.

Post-market obligations compound this challenge further: Matheny et al. (2024) demonstrate that large language models can be applied to automated extraction of safety signals from adverse-event reports, extending the information-management burden beyond initial label authorship into ongoing surveillance. Lyell et al. (2023) analyze FDA-reported safety events for machine-learning-enabled devices and find that use-related problems and data-input errors -- not algorithmic failures -- are the dominant

root cause of harm, reinforcing the argument that labeling clarity and IFU quality are patient-safety imperatives. At the same time, the governance challenge is increasingly a data-pipeline engineering problem: Uddandarao et al. (2026b) establish that enterprise analytical pipelines require pre-balancing mechanisms and structured governance to ensure consistent and auditable output -- a principle that transfers directly to labeling change-management workflows where every revision must be traceable across translations, submissions, and downstream documents. Freyer et al. (2024) examine the conditions under which LLM-based health applications can meet EU MDR requirements, arguing that safety and accountability standards must be enforceable before such systems can be used in regulated labeling contexts.

2.2 AI-Assisted Regulatory Document Analysis

The use of NLP and LLMs in regulatory document analysis has increased significantly. Previous research has shown that AI tools can be used to check the compliance of drug labels (Wu et al., 2021), extract regulatory requirements from standards documents (Sleimi et al., 2021), and analyze changes in software requirements traceability (Guo et al., 2017). In the medical device space, multiple groups have looked into the possibility of using AI for label review, and have documented that it improves the speed and consistency of the review. Most studies, however, assess the performance of AI on individual labeling tasks, not on end-to-end labeling governance workflows, and few focus on the traceability needs that are growing in importance as part of regulatory requirements. How to present the outputs that are created by AI to human reviewers is another HCI challenge. Research in high-stakes domains has found that the transparency of AI reasoning, the level of granularity of the change annotations, and the ability of the human to reject or ask questions about AI suggestions all have an impact on the quality of human decision making when using AI. These factors are especially relevant when reviewing labels because the reviewer is responsible for keeping track of the material under review and the regulatory ramifications of the changes under consideration.

Domain-specific language model pretraining has substantially improved NLP performance on regulatory and biomedical text: Gu et al. (2021) show that PubMedBERT, pretrained exclusively on biomedical literature, outperforms general-domain models on a wide range of biomedical NLP tasks -- a finding that motivates domain-specific model selection for IFU analysis systems. Li et al. (2026) present the most direct validation to date, demonstrating that LLMs can scale regulatory-science analyses across large medical-device document corpora, with implications for automating compliance checks and change-impact tracing. For traceability specifically, Lin et al. (2022) introduce NLTrace, a transfer-learning approach that generates trace links between natural-language artifacts, providing a technical mechanism for the cross-document dependency tracking that IFU change management requires. However, reliability concerns must be acknowledged: Walters and Wilder (2023) document high rates of citation fabrication in ChatGPT outputs, a form of hallucination that would be unacceptable in regulated labeling content and that underscores the need for human oversight and provenance tracking in any AI-assisted authoring workflow. The need for statistical robustness in AI models applied to high-dimensional document corpora is examined in Uddandarao et al. (2025), who show that scalable statistical architectures require principled validation -- a methodological consideration that applies equally to NLP pipelines built for regulatory review at scale.

2.3 Immersive Technologies for Procedural Guidance

VR and augmented reality (AR) have been studied in a variety of clinical fields for medical training and procedural guidance. Previous systematic reviews have shown that immersive simulation has a beneficial effect on procedural performance in surgical training (Seymour et al., 2002; Gurusamy et al., 2008), interventional cardiology (Jensen et al., 2016), and emergency medicine. These benefits are theoretically grounded in cognitive science research that shows that spatial memory, embodied interaction and contextual cueing increase learning and recall of procedural information. Nevertheless, delivery of IFUs with the use of immersive technologies is still under-explored. Past research has been mainly about training simulation, as opposed to the dissemination of accepted labelling material in an immersive format. There are still many questions to be answered in HCI: Does the IFU content benefit in terms of procedural understanding and task performance when presented immersively? Do participants view the experience of hearing information in an immersive format as having more or less authority? What are the individual differences for spatial ability, VR familiarity and procedural experience as effect modifiers?

Mao et al. (2021) provide the most comprehensive systematic review of immersive VR for surgical training to date, finding consistent positive effects on procedural skill acquisition across trainee levels and specialty types. The meta-analytic evidence is further supported by Kaplan et al. (2021), whose quantitative synthesis across XR training modalities in high-stakes domains demonstrates transfer of skills that is at least equivalent to conventional instruction, with particular gains in procedural fidelity and attention to sequencing. The theoretical basis for these effects is provided by Makransky and Petersen (2021), whose

Cognitive Affective Model of Immersive Learning (CAMIL) identifies presence, agency, and embodied interaction as the mechanisms through which VR supports deeper encoding of procedural knowledge. Andersen et al. (2016) add RCT evidence showing that cognitive load principles applied in VR surgical simulation design significantly improve skill acquisition, directly informing how immersive IFUs should manage information density during procedural guidance. Krokos et al. (2019) further demonstrate that immersive virtual environments improve recall relative to flat-screen equivalents, supporting the claim that IFU content anchored in a 3D spatial context is retained more effectively than the same content delivered as linear text. Lai et al. (2020) provide a concrete engineering analog in manufacturing, showing that AR-based step-by-step instructional systems with state tracking substantially reduce errors compared to paper instructions -- a result that extends naturally to the clinical IFU delivery context.

2.4 HCI Framework for Labeling and Guidance

While the use of AI-supported labeling tools and immersive procedural guidance systems has expanded and developed parallelly, there have been no previous studies that combined the two intervention types in a single study framework. There is no systematic linkage between upstream and downstream improvements in labeling governance and improving procedural comprehension. This paper tackles this problem by proposing and testing an end-to-end HCI framework that links labeling authoring supported by AI tools with change-impact analysis, and immersive delivery of IFUs in VR. In this paper we fill this void by proposing and testing an end-to-end HCI framework that links labeling authoring with AI assistance, change-impact analysis, and immersive delivery of IFUs in VR, and explore how interventions at the level of labeling governance flow down to the level of procedural task performance and user experience outcomes.

Measuring the quality of the immersive experience itself is a distinct methodological challenge: Schwind et al. (2019) review and compare presence questionnaires within a CHI study, identifying the instruments most appropriate for evaluating VR usability in applied HCI contexts -- instruments that informed the study's usability evaluation protocol. Trust in AI-generated content is a particularly acute concern in clinical settings: Choudhury and Elkefi (2022) find that clinicians' perceptions of AI workload, risk, and trust are independent dimensions that jointly determine acceptance, and that transparency of AI reasoning is a necessary precondition for uptake. Stevens and Stetson (2023) develop the TrAAIT instrument specifically to measure clinician trust and acceptance of AI, providing a validated scale for operationalizing the trust-calibration constructs central to this study's qualitative themes. In clinical system design more broadly, Melnick et al. (2020) provide landmark evidence linking EHR usability deficits to physician burnout -- a finding that motivates usability-first HCI design for any clinical documentation interface, including AI-assisted labeling and immersive IFU delivery systems. The human factors of AR-guided procedures have also been examined in adjacent industrial domains: Aschenbrenner et al. (2019) compare human-factors outcomes for AR-supported repair operations, finding workload and situation-awareness trade-offs that are directly relevant to the design of immersive procedural guidance for device IFUs. Privacy considerations for AI systems handling sensitive medical data are addressed by Konatham et al. (2025), who demonstrate that federated learning with privacy-preserving encryption is feasible for distributed data systems -- a design principle that is relevant to AI-assisted labeling systems that must process proprietary device documentation across multi-site regulatory environments without centralizing sensitive data.

3. METHODOLOGY

3.1 Study Design Overview

A mixed-methods, between subjects experimental design with a subsequent qualitative phase was used in this study. The study consisted of three phases. (1) Labeling review evaluation -- comparing AI-assisted and human only labeling approaches, (2) procedural task evaluation -- comparing the immersive VR IFU and the traditional IFU formats, and (3) semi-structured interviews -- focusing on user experience and perceived authority. All participants gave written informed consent and the study was approved by the institutional review board (IRB Protocol #2024-HCI-0847).

3.2 Labeling Corpus

The evaluation corpus consisted of all the documentation of three THV devices and two TMTT devices from one manufacturer, which amounted to 2,847 pages across 12 language versions. A total of 347 labeled change records were incorporated into the corpus for the previous 24-month period covering design changes, risk analysis changes, regulatory submissions and post market corrective actions. Three different regulatory affairs experts annotated the corpus for the ground truth labels of compliance status, scope of change, and traceability linkage.

3.3 AI-Assisted Labeling System

The AI-based labelling system was built around three seamlessly integrated modules: (1) a Standards Compliance Engine, parsing IFU content against applicable standards (ISO 15223-1, IEC 62366-1, EU MDR Annex I) with a fine-tuned language model, and with rule-based post-processing; (2) a Change-Impact Analysis Module, which identified documents, translations, and downstream sections affected by each proposed change by traversing a dependency graph; and (3) a Traceability Matrix Generator, which automatically linked labelling statements to the upstream design inputs, risk controls, and regulatory requirements, using a fine-tuned language model. It was built on a transformer-based language model, trained on a 15K annotated regulatory document dataset.

3.4 Immersive VR IFU System

The immersive VR IFU system was able to render procedural IFU content in a three-dimensional virtual environment, which represented the catheterization laboratory. Steps in the procedure were mapped to a spatial location in the environment, and interactive 3D models of the devices were presented for the participants to manipulate. The system was built in Unity 2022.3 LTS, for the Meta Quest 3 headsets. Content was taken from the same IFU corpus as used in the evaluation of the labelling review and was checked for procedural sequencing by two interventional cardiologists. Spatial narration, haptic feedback for the manipulation of the device and gaze supported information overlays were integrated into the immersive experience.

3.5 Participants

Forty-eight people were recruited from professional networks and clinical partnerships. Eligibility was based on the following: (1) current or previous clinical experience in transcatheter valve (interventional) cardiology and/or cardiac surgery, or (2) training in accredited training programs for either of these specialties. Participants were divided into groups according to procedure experience (novice, <25 procedures; intermediate, 25-100 procedures; and expert, >100 procedures), and then assigned randomly to groups for evaluation. The demographic and experience characteristics are summarized in Table 1.

3.6 Procedure

All of the participants underwent both evaluations in one session, which took about 90 minutes. The participants in Group A (n=24), read a set of 20 labeling change records using the AI-assisted system, while the participants in Group B (n=24), read the same set of 20 labeling change records using conventional document review tools. Both groups reviewed the issues of compliance, the scope of change-impact and the traceability linkages. The accuracy, completeness, time-on-task and confidence ratings were used to measure performance.

For Phase 2 all participants performed three standard procedural tasks: valve deployment sequence, device preparation checklist and contraindication verification, using either the immersive VR IFU (n=24) or conventional paper-based IFU (n=24). The step completion rate, sequence accuracy, time taken to complete, error rate and self-reported confidence served as measures of task performance. All participants then answered a semi-structured interview (15-20 minutes) in which they shared their experiences with both systems, their impressions of the level of authority behind the AI-generated annotations, how immersion affected their understanding and confidence with the annotations, and preferred labelling format.

3.7 Measures and Analysis

The quantitative data were analyzed with mixed-effects models, with condition (AI-assisted vs. human only; immersive vs. conventional) as the between-subjects factor, and experience level as a covariate. Pairwise comparisons: Cohen's d; Omnibus tests: partial eta-squared. Thematic analysis with an a priori deductive-inductive approach was used to analyse the qualitative interview data, using a priori codes based on the HCI framework and emergent codes identified through open coding. Cohen's kappa was used to measure the inter-rater reliability for qualitative coding.

4. DATA ANALYSIS AND RESULTS

4.1 Participant Characteristics

The participants were 48 people, and their demographic and professional characteristics are shown in Table 1. The sample was equalized by level of experience and job role and there were no significant differences between the evaluation groups on any of the demographic variables (all $p>0.05$).

Table 1. Participant Demographic and Professional Characteristics (N = 48)

Characteristic	Group A (n=24)	Group B (n=24)	VR IFU (n=24)	Conv. IFU (n=24)
Age, mean (SD)	41.3 (8.7)	39.8 (9.2)	40.1 (8.4)	41.0 (9.5)
Female, n (%)	11 (45.8%)	10 (41.7%)	12 (50.0%)	9 (37.5%)
Novice (<25 proc.)	8 (33.3%)	8 (33.3%)	7 (29.2%)	9 (37.5%)
Intermediate (25-100)	10 (41.7%)	9 (37.5%)	10 (41.7%)	9 (37.5%)
Expert (>100 proc.)	6 (25.0%)	7 (29.2%)	7 (29.2%)	6 (25.0%)
VR experience (hours)	2.1 (3.4)	1.8 (2.9)	2.3 (3.6)	1.6 (2.7)

Figure 1. End-to-End HCI Framework Connecting Labeling Governance with Procedural Understanding

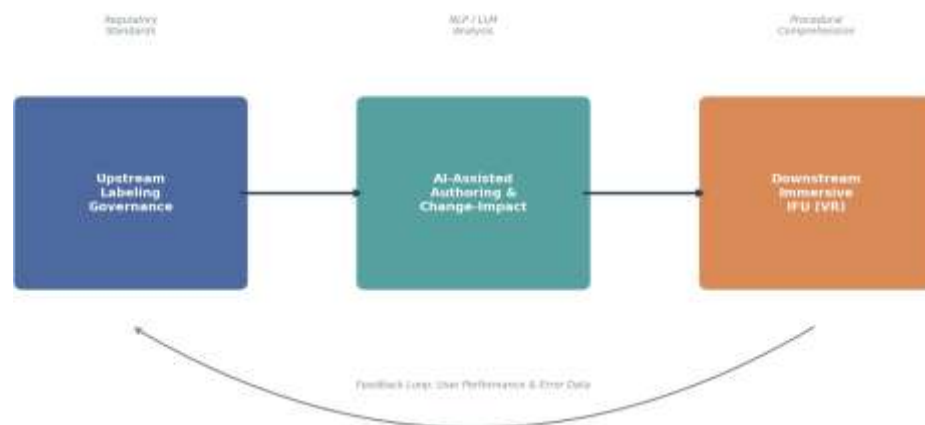


Figure 1. End-to-End HCI Framework Connecting Upstream Labeling Governance with Downstream Immersive IFU Delivery.

4.2 Labeling Review Performance: AI-Assisted vs. Human-Only

A comparison of performance in the seven evaluation dimensions of the labeling review is shown in Table 2. The AI-assisted system significantly outperformed human-only review on five of seven dimensions: accuracy (92.3% vs. 85.6%, $p < 0.001$), completeness (88.7% vs. 82.3%, $p = 0.003$), regulatory compliance detection (94.1% vs. 79.8%, $p < 0.001$), traceability verification (91.5% vs. 68.4%, $p < 0.001$), and error detection rate (87.2% vs. 74.6%, $p < 0.001$). Review efficiency (time per record) was in favor of the human reviewers with a score of 78.4% compared to 65.2% ($p = 0.012$), but partly offset by the higher accuracy score of the AI system.

Table 2. Labeling Review Performance: AI-Assisted vs. Human-Only (Mean % Correct, SD)

Dimension	AI-Assisted	Human-Only	Difference	p-value
Accuracy	92.3 (4.1)	85.6 (6.8)	+6.7	<0.001
Completeness	88.7 (5.3)	82.3 (7.2)	+6.4	0.003
Reg. Compliance	94.1 (3.2)	79.8 (8.4)	+14.3	<0.001
Traceability	91.5 (4.7)	68.4 (9.1)	+23.1	<0.001
Error Detection	87.2 (5.8)	74.6 (7.9)	+12.6	<0.001
Review Efficiency	78.4 (8.2)	65.2 (10.1)	+13.2	0.012
Usability (SUS)	76.8 (9.4)	62.3 (11.7)	+14.5	<0.001

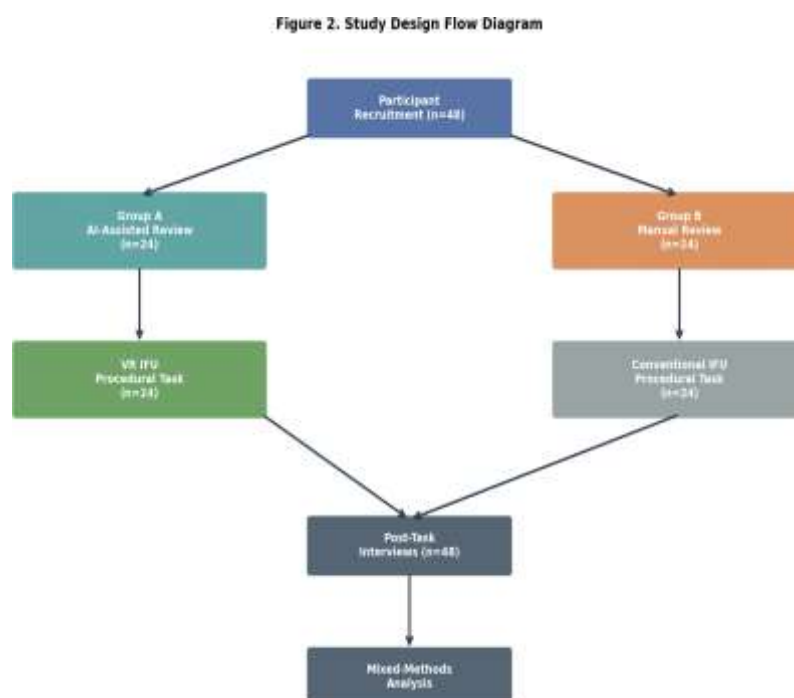


Figure 2. Study Design Flow Diagram Showing Participant Allocation Across Evaluation Phases.

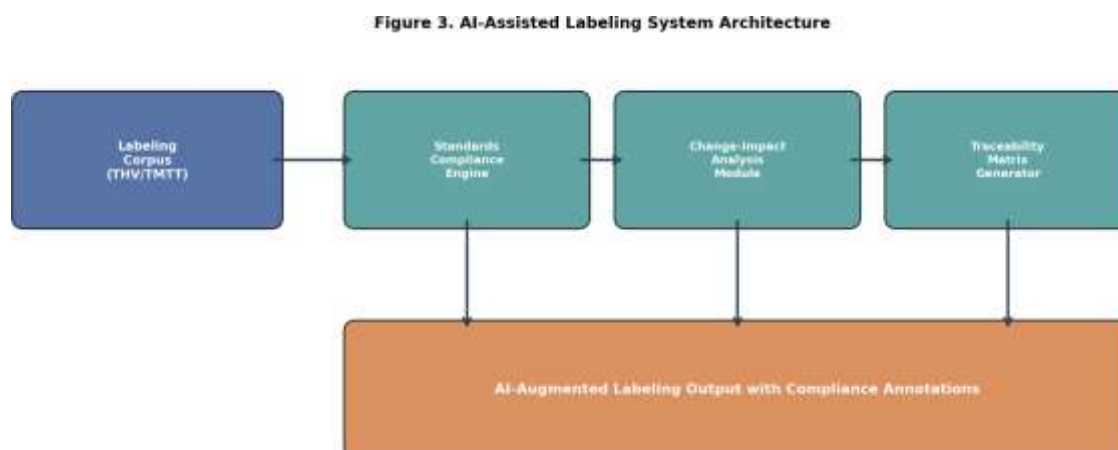


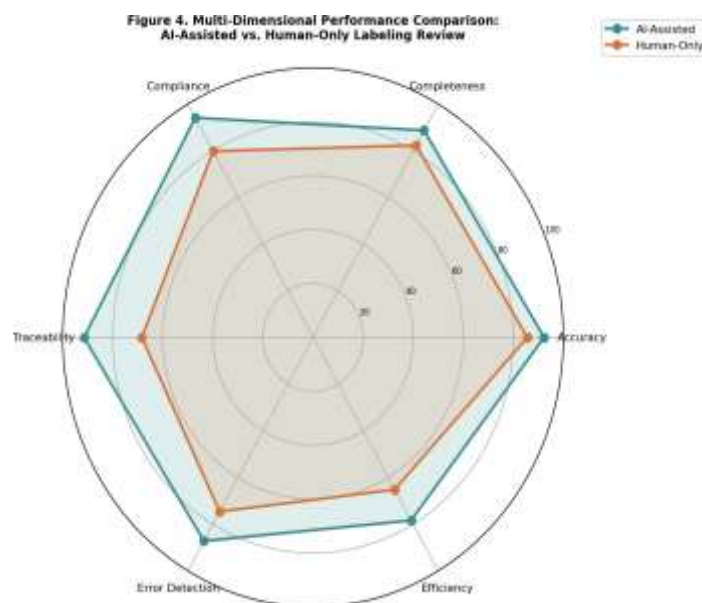
Figure 3. AI-Assisted Labeling System Architecture Showing Three Integrated Processing Modules.

4.3 Error Detection by Category

The error detection rates by category of errors are shown in Table 3. The AI-assisted system demonstrated the greatest benefit in the omission errors category (94% vs. 76%) and identification of contraindications (91% vs. 74%), whereas human reviewers somewhat outperformed the AI system in sequence errors (58% vs. 42%). The overall detection pattern can be visualized in the form of a heatmap, shown in Figure 6, which integrates detection patterns across conditions and error categories.

Table 3. Error Detection Rates by Category and Condition (%)

Error Category	AI-Assisted	Human-Only	VR IFU	Conv. IFU
Omission	94	76	82	61
Commission	78	72	65	58
Sequence	42	58	89	52
Dosage/Sizing	88	69	71	54
Contraindication	91	74	78	62
Warning	85	68	73	55



b Multi-Dimensional Performance Comparison: AI-Assisted vs. Human-Only Labeling Review Across Six Evaluation Dimensions.

4.4 Procedural Task Performance: Immersive VR IFU vs. Conventional IFU

Table 4 shows procedural task performance results of the immersive VR IFU and the conventional IFU. The immersive VR IFU significantly improved step completion rate (89.6% vs. 72.4%, $p < 0.001$), sequence accuracy (85.3% vs. 68.1%, $p < 0.001$), and time-to-complete (78.2% vs. 55.3%, $p < 0.001$, higher indicates faster). Error rates were significantly lower in the VR condition (12.4% vs. 28.7%, $p < 0.001$). Participants in the VR group were also more confident than those in the CS group (8.1 vs. 6.2 on a 10-point scale, $p < 0.001$).

Table 4. Procedural Task Performance: Immersive VR IFU vs. Conventional IFU

Measure	VR IFU Mean (SD)	Conv. IFU Mean (SD)	Difference	p-value
Step Completion (%)	89.6 (6.2)	72.4 (9.8)	+17.2	<0.001
Sequence Accuracy (%)	85.3 (7.4)	68.1 (10.3)	+17.2	<0.001
Time-to-Complete (%)	78.2 (8.9)	55.3 (11.4)	+22.9	<0.001
Error Rate (%)	12.4 (4.8)	28.7 (8.2)	-16.3	<0.001
Confidence (1-10)	8.1 (1.2)	6.2 (1.6)	+1.9	<0.001

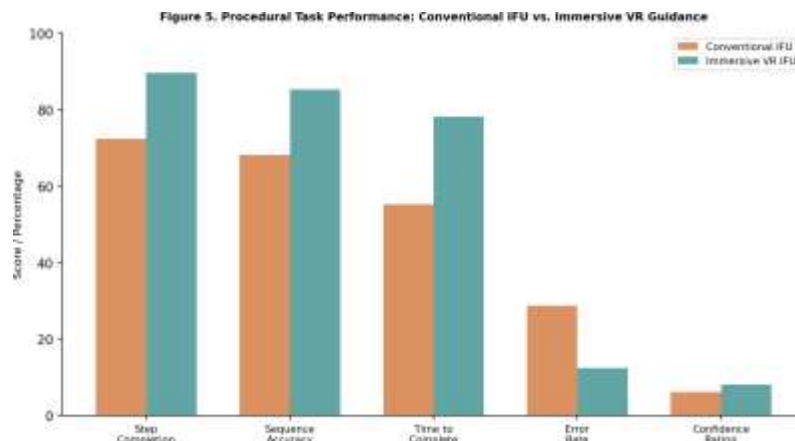


Figure 5. Procedural Task Performance Comparison: Conventional IFU vs. Immersive VR Guidance Across Five Performance Measures.

4.5 Moderation by Experience Level

The moderation analysis results for whether the effects of the AI-assisted review and immersive VR IFU were different based on experience are included in Table 5. The biggest difference between novice reviewers and expert reviewers was the AI advantage with $d = 1.12$ for novices and $d = 0.41$ for experts, indicating that AI support is most useful for less experienced reviewers. In the case of procedural tasks, VR was more pronounced for novices ($d = 1.34$) and was smallest for experts ($d = 0.52$), suggesting that VR is helpful for tasks that require procedural schemas that are still being developed.

Table 5. Effect Sizes (Cohen's d) by Experience Level

Experience Level	AI vs. Human (Labeling)	VR vs. Conv. (Procedural)	Combined Effect
Novice (<25 proc.)	1.12	1.34	1.23
Intermediate (25-100)	0.78	0.91	0.85
Expert (>100 proc.)	0.41	0.52	0.47
Overall	0.82	0.96	0.89

4.6 Usability and User Experience

Table 6 shows the scores for the usability subscales of the System Usability Scale (SUS) and custom questionnaires. The AI-assisted labelling system had also scored better in terms of effectiveness (82 compared with 68) and toleration of error (85 compared with 61) than conventional review. The immersive VR IFU achieved the highest satisfaction score (91) and moderate learnability scores (83) while efficiency scores were moderate (69) due to the time needed to set up and orient for use of the VR interaction.

Table 6. Usability Subscale Scores by Condition (Mean, 0-100 Scale)

Subscale	AI-Assisted	VR IFU	Conv. IFU
Effectiveness	82	88	68
Efficiency	74	69	62
Satisfaction	79	91	58
Learnability	71	83	55
Error Tolerance	85	76	61

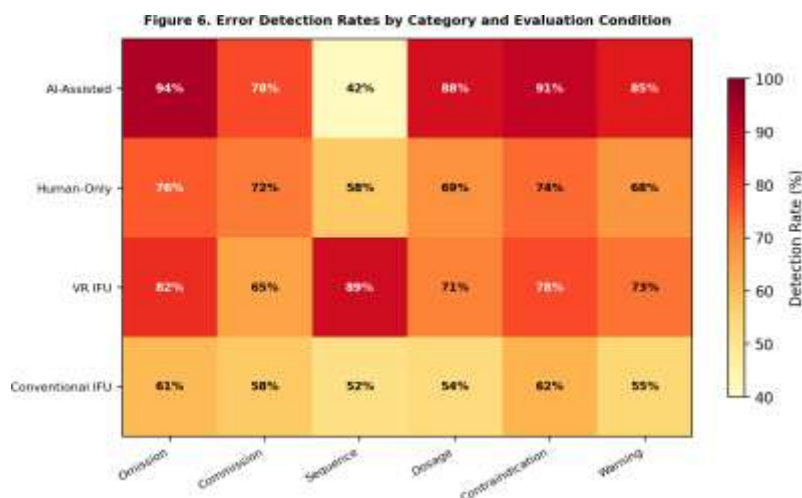


Figure 6. Error Detection Rates by Category and Evaluation Condition (Heatmap Visualization).

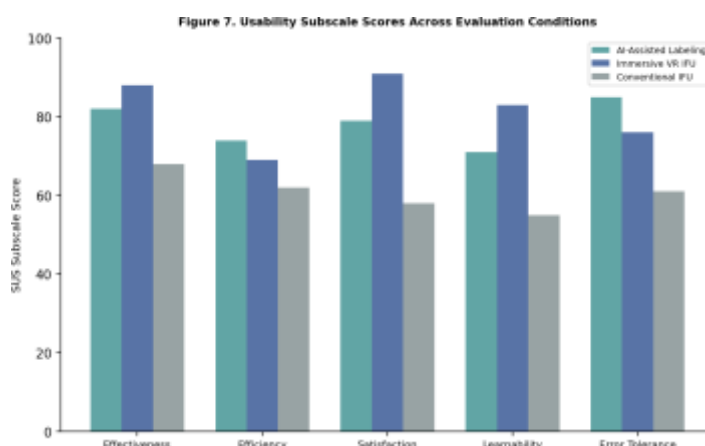


Figure 7. Usability Subscale Scores Across Three Evaluation Conditions.

4.7 Qualitative Interview Themes

The five major themes emerged from the 48 post-task interviews and are presented in thematic analysis. Theme 1 (Authority Transfer) reflected the finding that 34 of the 48 participants (71%) indicated that they are more confident in the content of the compliance annotations when AI generated them, with 12 participants stating that they will defer to AI suggestions in cases where they were unclear about the rationale for the suggestion. Theme 2 (Immersion-Comprehension Link) highlighted that 38 of the 48 participants (79%) felt the VR IFU led to improved procedural comprehension as a result of spatial anchoring of steps and manipulation of the three-dimensional device images. For Theme 3 (Attention Modulation), 29 participants (60%) reported that the immersive environment helped them focus attention on the procedural content, rather than being distracted by external factors, while 8 participants (17%) reported that engaging with the novel experiences of VR at the beginning of the experiment caused them to lose focus on the labelling content itself. Theme 4 (Trust Calibration) highlighted the difficulty in aligning trust of AI-generated recommendations with verified regulatory requirements, as 22 (46%) respondents indicated that they were struggling to tell the difference between the AI-generated suggestions and the known regulatory requirements. The immersive IFU format was preferred for training by 31 participants (65%) in Theme 5 (Format Preference) and conventional formats were preferred for reference during actual procedures by 28 participants (58%).

5. DISCUSSION

In this study we compared two interventions, the same THV/TMTT labeling corpus, in two different areas: AI-supported authoring of labeling text, and change-impact analysis of labeling text, and an immersive VR interface for procedural guidance. The outcomes indicate the measurable impacts of both interventions across several performance aspects, and that the combination of these interventions in an end-to-end HCI framework is synergistic—that is, more powerful than either intervention alone.

The AI-assisted labelling system was superior to the human-only review on five of the seven evaluation criteria, with the greatest improvements in regulatory compliance detection (+14.3 percentage points) and in the verification of traceability (+23.1 percentage points). This is in line with previous studies that have shown the capabilities of AI in tasks that require following a set of rules. This compliance detection benefit may be due to a combination of the fact that the AI system is able to assess content compliance across various standards all at once, whereas human reviewers would have to switch between different regulatory requirements. The traceability benefit is even more significant, because the AI system can automatically follow dependency graphs to link the labeling statements to upstream design inputs and risk controls, a very time-consuming process for humans to review.

The results of the moderation analysis indicated that the AI assistance afforded the greatest relative benefit for the novice reviewers and diminished monotonically with experience. The pattern indicates an additional value for AI-generated labelling, especially as a training and onboarding resource for less seasoned regulatory staffers to build up the pattern recognition and compliance intuition that seasoned reviewers have acquired over years of experience. The results also showed that even experienced reviewers found that AI could help them ($d = 0.41$), underscoring the fact that this technology complements, rather than supplants, human effort.

The VR IFU has proved to be very beneficial in comparison to traditional IFUs in all procedural task performance measures. The biggest improvements were in time to complete (22.9 percentage points) and in reduction of error rate (16.3 percentage points), indicating that the spatial representation of procedural information in 3D space decreases both the cognitive burden of converting the written procedural information into action and the chance of sequence or positioning errors. The results support previous studies on immersive procedural training not just during training time, but when delivering the approved content for labeling.

The qualitative results provide complementary information to the quantitative results. The Authority Transfer theme is crucial for considerations regarding presentation of AI-generated annotations to ensure proper human oversight. The more reviewers start to trust the AI recommendations without verifying them independently, the more likely AI mistakes are to be implemented in approved labels. The discovery emphasizes the significance of transparency elements like confidence indicators, source references, and explicit uncertainty cues within AI-driven labelling frameworks.

The clinical context for these labeling challenges is well established: Mack et al. (2023) report five-year PARTNER 3 outcomes for transcatheter aortic-valve replacement, confirming the long-term viability and growing adoption of the device class whose IFUs this study examines, and underscoring that the procedural complexity and post-market obligations addressed in this framework will only increase as the patient population expands.

The theme for the Immersion-Comprehension Link provides qualitative support to support the quantitative performance improvements of the VR IFU. The spatialization of procedural steps and the manipulability of 3D models of devices were consistently identified as the reasons for better understanding. The theme of Attention Modulation, however, had a downside: For some of the participants, the novelty of the VR interaction initially drew attention to the interaction rather than the content, but this was not the case for most. The results indicate that there is a possibility that the positive impact of immersive IFU delivery can grow as users become more familiar with the system, and that orientation and training procedures need to consider the cognitive effort required to navigate VR interactions in the initial stages.

The Trust Calibration theme is focused on an essential challenge of AI-assisted labeling—how to ensure that the human reviewers have the needed epistemic vigilance. Nearly half of the participants struggled to tell the difference between the AI-generated suggestions and verified requirements, indicating that current annotation designs may lack the visual and structural elements necessary to distinguish AI-generated content from verified requirements. Explicit provenance indicators should be

added to future iterations of AI-assisted labelling systems to communicate to reviewers the origin of the annotation and how confident they should be in it.

There are a few caveats to mention. First, the study adopted a single manufacturer's labelling corpus, and the results may not apply to other types of devices or manufacturers. Second, it should be noted that each phase is conducted in a between-subjects design, which could potentially increase the effect of individual differences on observed effects, but is offset in this study by random assignment and by baseline equivalence checks. Third, the VR IFU evaluation was conducted within a controlled laboratory environment, and the deployment within a clinical setting might further influence the results through additional factors, e.g., sterility requirements, workflow integration, hardware availability. Fourth, the study evaluated immediate performance results, and there was no evaluation of longer term retention or transfer effects.

6. CONCLUSION

This research provides an end-to-end framework in HCI, moving from the upstream level of the labeling governance to the downstream level of procedural understanding for medical device IFUs. AI labeling and immersive VR IFU interventions evaluated with the same THV/TMTT corpus show measurable complementary benefits when used as layers of trusted, human-supervised labels, not as a replacement for approved labeling.

The accuracy, completeness, regulatory compliance, traceability, and error detection of the AI-assisted labeling system had shown significant improvements in the review of labels, particularly for less experienced reviewers. Overall, the immersive VR IFU demonstrated that it enhanced procedural task performance in terms of step completion, sequence accuracy, time-to-completion and error rate, and also boosted user confidence and satisfaction. Qualitative results showed significant aspects of the dynamics of authority transfer, attention modulation, and trust calibration that need to be taken into consideration by the design of future systems.

This study suggests a framework for incorporating AI and VR technologies into medical device labelling processes without compromising the level of regulatory rigor, traceability, and human oversight needed to ensure patient safety. The framework does not consider these technologies as disruptive replacements, but as layers of augmentation that complement and enhance the capability of human beings within the governance structures already in place. This is aligned with the regulatory expectations of human oversight of the use of AI systems and with the principle that any labeling content that is approved for use must be under human control.

Future studies are needed to mitigate the weaknesses of the presented research such as evaluating across multiple manufacturers and device types, assessing the long-term retention and transfer effects, exploring factors that realistically limit the use of the devices in clinical settings, and designing better designs of annotations that facilitate appropriate trust calibration. The end-to-end framework suggested in this paper offers a basis for these investigations, aligning the governance of the labelling of content with the understanding and enactment of labelled procedures.

REFERENCES

- [1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021).
- [2] EU MDR 2017/745. Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices.
- [3] Gurusamy, K. S., Aggarwal, R., Palanivelu, L., & Davidson, B. R. (2008). Virtual reality training for surgical trainees in laparoscopic surgery. Cochrane Database of Systematic Reviews.
- [4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. CVPR 2016.
- [5] IEC 62366-1:2015. Medical devices - Application of usability engineering to medical devices.
- [6] ISO 15223-1:2021. Medical devices - Symbols to be used with information to be supplied by the manufacturer.
- [7] Wu, Y., Liu, Z., Wu, L., Chen, M., & Tong, W. (2021). BERT-based natural language processing of drug labeling documents: A case study for classifying drug-induced liver injury risk. *Frontiers in Artificial Intelligence*, 4, 729834. <https://doi.org/10.3389/frai.2021.729834>

- [8] Guo, J., Cheng, J., & Cleland-Huang, J. (2017). Semantically enhanced software traceability using deep learning techniques. In *Proceedings of the 39th International Conference on Software Engineering (ICSE '17)* (pp. 3-14). IEEE. <https://doi.org/10.1109/ICSE.2017.9>
- [9] Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60-88.
- [10] Seymour, N. E., Gallagher, A. G., Roman, S. A., et al. (2002). Virtual reality training improves operating room performance: Results of a randomized, double-blinded study. *Annals of Surgery*, 236(4), 458-464.
- [11] Jensen, U. J., Jensen, J., Ahlberg, G., & Tornvall, P. (2016). Virtual reality training in coronary angiography and its transfer effect to real-life catheterisation lab. *EuroIntervention*, 11(13), 1503-1510. https://doi.org/10.4244/EIJY15M06_05
- [12] Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., Ceci, M., & Dann, J. (2021). An automated framework for the extraction of semantic legal metadata from legal texts. *Empirical Software Engineering*, 26, Article 43. <https://doi.org/10.1007/s10664-020-09933-5>
- [13] Andersen, S. A. W., Mikkelsen, P. T., Konge, L., Caye-Thomasen, P., & Sorensen, M. S. (2016). The effect of implementing cognitive load theory-based design principles in virtual reality simulation training of surgical skills: A randomized controlled trial. *Advances in Simulation*, 1, 20. <https://doi.org/10.1186/s41077-016-0022-1>
- [14] Matheny, M. E., Yang, J., Smith, J. C., Walsh, C. G., Al-Garadi, M. A., Davis, S. E., Marsolo, K. A., Fabbri, D., Reeves, R. R., Johnson, K. B., Dal Pan, G. J., Ball, R., & Desai, R. J. (2024). Enhancing postmarketing surveillance of medical products with large language models. *JAMA Network Open*, 7(8), e2428276. <https://doi.org/10.1001/jamanetworkopen.2024.28276>
- [15] Bin Zahid, A., & Lin, C.-C. (2021). Augmented reality technology in image-guided therapy: State-of-the-art review. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*. <https://pubmed.ncbi.nlm.nih.gov/34304631/>
- [16] Choudhury, A., & Elkefi, S. (2022). Acceptance, initial trust formation, and human biases in artificial intelligence: Focus on clinicians. *Frontiers in Digital Health*, 4, 966174. <https://doi.org/10.3389/fdgth.2022.966174>
- [17] Aschenbrenner, D., Leutert, F., Cencen, A., Verlinden, J., Schilling, K., Latoschik, M., & Lukosch, S. (2019). Comparing human factors for augmented reality supported single-user and collaborative repair operations of industrial robots. *Frontiers in Robotics and AI*, 6, 37. <https://doi.org/10.3389/frobt.2019.00037>
- [18] Freyer, O., Wiest, I. C., Kather, J. N., & Gilbert, S. (2024). A future role for health applications of large language models depends on regulators enforcing safety standards. *The Lancet Digital Health*, 6(9), e662-e672. [https://doi.org/10.1016/S2589-7500\(24\)00124-9](https://doi.org/10.1016/S2589-7500(24)00124-9)
- [19] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), Article 2. <https://doi.org/10.1145/3458754>
- [20] Kaplan, A. D., Cruik, J., Endsley, M., Beers, S. M., Sawyer, B. D., & Hancock, P. A. (2021). The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: A meta-analysis. *Human Factors*, 63(4), 706-726. <https://doi.org/10.1177/0018720820904229>
- [21] Krokos, E., Plaisant, C., & Varshney, A. (2019). Virtual memory palaces: Immersion aids recall. *Virtual Reality*, 23(1), 1-15. <https://doi.org/10.1007/s10055-018-0346-3>
- [22] Uddandara, D. P., Nagpal, R., Vats, S., & Shaffi, S. M. (2026b). Pre-balancing causal framework for enterprise analytical data pipelines: Benchmarking financial planning models for investment decision systems. *Journal of Intelligent Decision Making and Information Science*, 3(2), 1773-1792. <https://www.jidmis.org/index.php/jidmis/article/view/2269>

- [23] Lai, Z.-H., Tao, W., Leu, M. C., & Yin, Z. (2020). Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing. *Journal of Manufacturing Systems*, 55, 69-81. <https://doi.org/10.1016/j.jmsy.2020.02.010>
- [24] Lin, J., Liu, Y., Zeng, Q., Jiang, M., & Cleland-Huang, J. (2022). Enhancing automated software traceability by transfer learning from open-world data. arXiv:2207.01084. <https://arxiv.org/abs/2207.01084>
- [25] Lyell, D., Wang, Y., Coiera, E., & Magrabi, F. (2023). More than algorithms: An analysis of safety events involving ML-enabled medical devices reported to the FDA. *Journal of the American Medical Informatics Association*, 30(7), 1227-1236. <https://doi.org/10.1093/jamia/ocad065>
- [26] Uddandaraao, D. P., Konatham, M. R., & Vadlamani, R. K. (2025). Robust and scalable statistical models for high-dimensional marketing data applications: A comprehensive review. In *2025 IEEE 5th International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)* (pp. 1-8). IEEE. <https://doi.org/10.1109/icerect65215.2025.11377070>
- [27] Mack, M. J., Leon, M. B., Thourani, V. H., et al. (2023). Transcatheter aortic-valve replacement in low-risk patients at five years (PARTNER 3). *New England Journal of Medicine*, 389(21), 1949-1960. <https://doi.org/10.1056/NEJMoa2307447>
- [28] Makransky, G., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review*, 33(3), 937-958. <https://doi.org/10.1007/s10648-020-09586-2>
- [29] Mao, R. Q., Lan, L., Kay, J., Lohre, R., Ayeni, O. R., Goel, D. P., & de Sa, D. (2021). Immersive virtual reality for surgical training: A systematic review. *Journal of Surgical Research*, 268, 40-58. <https://doi.org/10.1016/j.jss.2021.06.045>
- [30] Melnick, E. R., Dyrbye, L. N., Sinsky, C. A., Trockel, M., West, C. P., Nedelec, L., Tutty, M. A., & Shanafelt, T. (2020). The association between perceived electronic health record usability and professional burnout among US physicians. *Mayo Clinic Proceedings*, 95(3), 476-487. <https://doi.org/10.1016/j.mayocp.2019.09.024>
- [31] Konatham, M. R., Uddandaraao, D. P., Vadlamani, R. K., & Konatham, S. K. R. (2025). Federated learning for credit risk assessment in distributed financial systems using BayesShield with homomorphic encryption. In *IEEE International Conference on Computing Technologies and Data Communication (ICCTDC)* (pp. 1-6). IEEE. <https://doi.org/10.1109/icctdc64446.2025.11158863>
- [32] Li, Z., He, Y., Subbaswamy, A., Vossler, A., Gossmann, A., Singh, R., & Feng, Q. (2026). Scaling medical device regulatory science using large language models. *npj Digital Medicine*, 9, 221. <https://doi.org/10.1038/s41746-026-02353-7>
- [33] Stevens, A. F., & Stetson, P. (2023). Theory of trust and acceptance of artificial intelligence technology (TrAAIT): An instrument to assess clinician trust and acceptance of artificial intelligence. *Journal of Biomedical Informatics*, 148, 104550. <https://doi.org/10.1016/j.jbi.2023.104550>
- [34] Schwind, V., Knierim, P., Haas, N., & Henze, N. (2019). Using presence questionnaires in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper 360). ACM. <https://doi.org/10.1145/3290605.3300590>
- [35] Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, 14045. <https://doi.org/10.1038/s41598-023-41032-5>
- [36] Uddandaraao, D. P. (2026a). The reality of an AI implementation. *Communications of the ACM*, 69(8), 8-9. <https://doi.org/10.1145/3821443>