# Human-AI Collaboration for 2D/3D Registration Quality Assurance

Sue Min Cho, Russell H. Taylor, and Mathias Unberath

Johns Hopkins University, Baltimore MD, USA
scho72@jhu.edu

**Abstract.** As surgery increasingly integrates advanced imaging, algorithms, and robotics to automate complex tasks, human judgment of system correctness remains a vital safeguard for patient safety. A critical example is 2D/3D registration, where even small misalignments can lead to surgical errors. Current visualization strategies alone are insufficient for reliable misalignment detection, highlighting the need for algorithmic decision-support. Using an AI framework for 2D/3D registration quality assessment, augmented with explainable AI (XAI) mechanisms to clarify model predictions, we conducted a user study (N=18) systematically comparing decision-making across three conditions: Human-only, Human+AI, and Human+XAI. We evaluated both objective performance measures (accuracy, sensitivity, precision, specificity) and subjective factors (workload, trust, understanding). Collaborative paradigms (Human+AI and Human+XAI) showed improved sensitivity, precision, and specificity compared to Human-only. Participants experienced significantly lower workload in collaborative conditions relative to the Human-only condition. Moreover, participants reported greater understanding of AI predictions in the Human+XAI condition than in Human+AI, although no significant differences were observed between the two collaborative paradigms in perceived trust or workload. Human-AI collaboration can enhance 2D/3D registration quality assurance, with explainability mechanisms improving user understanding. Future work should refine XAI designs to translate improved understanding into optimized decision-making performance.

**Keywords:** Assured autonomy · 2D/3D registration · image-guided surgery · explainable AI · human-centered AI · human-AI interaction.

## 1 Introduction

Surgery is undergoing a profound digital transformation, evolving from procedures performed solely by human experts to those guided by sophisticated imaging and advanced algorithms, assisted or automated through robotic technology. Yet, even as technology reshapes how modern surgery is delivered, human judgment remains indispensable for ensuring correct system function, and thus, patient safety [9]. The emerging concept of human-centered assurance underscores the need to integrate human operators into complex, technology-assisted

workflows [3, 1]. Still, the precise roles and responsibilities of these operators are far from clearly defined [4]. As surgical platforms become more automated, new tasks—such as verifying advanced algorithmic outputs—will increasingly fall to staff members or specialists who may not hold traditional clinical titles. Consequently, understanding how these "operators" perceive, interpret, and act on algorithmic information is essential for robust safety assurance.

This challenge is exemplified in the assurance of image-based navigation systems, where humans must verify the correctness of algorithmic alignments that directly affect surgical precision. In semi- and fully autonomous minimally invasive surgery, human oversight remains indispensable even as autonomy increases. Although machine intelligence promises to reduce errors and enhance precision, the responsibility for ensuring patient safety remains squarely with those in the operating room. Surgical robots perform both routine maneuvers, such as navigating through unoccupied space, and critical tasks, such as bone drilling or implant placement, where even millimeter-scale errors can cause permanent damage. These high-stakes "branching decisions" require not only accurate intraoperative guidance but also sufficient transparency for human operators to confidently validate algorithmic outputs.

These assurance challenges are vividly illustrated by 2D/3D registration, a key enabler in image-guided navigation that aligns intraoperative 2D fluoroscopic images with pre- or intraoperative 3D volumes. Indeed, image-based navigation—where the relative poses of surgical plans, instruments, and anatomy are estimated directly from image data—has long been heralded as the future of navigated surgery [8, 7]. However, despite significant gains in the accuracy and autonomy of 2D/3D registration through both optimization-based and deep learning approaches [11], ensuring quality and safety remains a central challenge. Because image-based navigation is most frequently used in delicate anatomy, such as the spine, even small misalignments (on the order of several millimeters) can lead to critical deviations in tool placement or implant positioning.

To address this gap, this work presents a human-factors evaluation of algorithmic and explainable AI support for quality assurance in image-guided surgery. Specifically, we: 1) employ a learning-based AI model for 2D/3D registration quality assessment, integrated with explainability mechanisms that offer transparent justifications for model predictions, 2) conduct a systematic user study (N=18) comparing three decision-making conditions: Human-only, Human+AI (AI assistance without explanations), and Human+XAI (AI assistance with explainability), and 3) evaluate both objective performance metrics and subjective user experience, quantifying how AI assistance and explainability affect human decision-making in safety-critical verification tasks.

## 2 Methods

### 2.1 AI Model for Registration Quality Assessment

We employ a learning-based model that classifies a 2D/3D registration result as successful or failed, where success is defined as mean target registration error

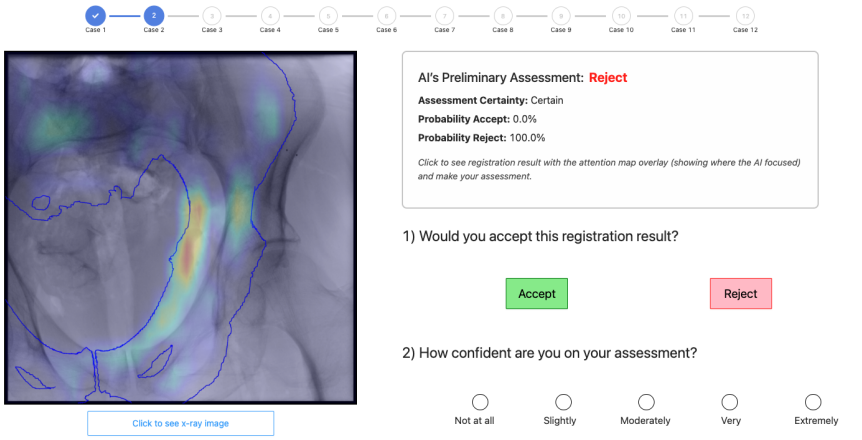


**Fig. 1.** User interface in the Human–XAI condition. Participants viewed the AI's classification decision ("Accept" or "Reject") along with confidence scores and Grad-CAM heatmaps overlaid on the X-ray images and registration overlay, highlighting spatial regions that influenced the AI's judgment.

(mTRE) below 2 mm following orthopedic surgical standards [6]. The input pair consists of the intraoperative fluoroscopic image and a digitally reconstructed radiograph (DRR) rendered at the estimated pose. The two images are concatenated channel-wise (early fusion) and passed through a convolutional backbone, after which spatial cross-attention is applied bidirectionally across two channel-wise feature groups to capture long-range dependencies indicative of global misalignment. A fully connected layer produces the binary decision. The model was trained on pelvic fluoroscopy data using leave-one-specimen-out cross-validation, achieving 80.1% accuracy. To support human oversight, we expose two explainability outputs: Grad-CAM [10] heatmaps computed at the last convolutional layer to localize image regions driving the prediction, and a conformal-prediction-based confidence indicator that flags cases of model uncertainty.

### 2.2 Human-AI User Study Design

18 participants (8 males, 10 females), all with a STEM background, were recruited for the user study. This demographic was chosen to represent potential "operators", such as clinical engineers, company specialists, and validation personnel, who are often involved in safety assurance and validation of computer-assisted intervention systems in clinical settings. The study design was approved by the local IRB, and informed consent was obtained from all participants prior to their involvement.

The user interface (Fig. 1) was implemented using a Next.js-based browser platform for data collection. At the start of the study, participants received

instructions and were presented with example cases demonstrating various registration offsets. The study used a within-subjects design where participants completed three conditions: Human-only (No AI), Human+AI (AI assistance without explanations), and Human+XAI (AI assistance with explainability). The Human-only condition was always presented first to ensure that participants established a baseline understanding of the task without AI influence. The order of the Human+AI and Human+XAI conditions was randomized across participants to counterbalance potential ordering effects. During each condition, participants completed 12 assessment tasks. For each task, an X-ray image and its registration overlay were shown, and participants were asked to assess the alignment as "Accept" or "Reject," and provide their confidence levels using a 5-point Likert scale, from "Not at all" to "Extremely."

In the Human+AI condition, participants were additionally shown the AI's classification decision (Accept or Reject). In the Human+XAI condition, participants were shown the AI's confidence scores for its decision and Grad-CAM heatmaps, which were overlayed on the X-ray images to highlight spatial regions influencing the AI's predictions (Fig. 1). The order of X-ray images and registration offsets was randomized for each participant using an automated randomization procedure integrated into the experimental platform. After completing each condition, participants completed a short survey evaluating perceived workload and their experience with the AI assistance, if given. A post-study survey gathered demographic information.

### 2.3 Evaluation

**Performance Metrics** Performance was quantitatively evaluated using standard classification metrics: *accuracy*, *sensitivity* (true positive rate, TPR), *precision*, and *specificity* (true negative rate, TNR). Participants judged registration quality by choosing either "Accept" or "Reject," which were compared against the ground truth. A correct "Accept" corresponds to a *true positive* (TP) and a correct "Reject" to a *true negative* (TN), while incorrect decisions constitute *false positives* (FP) and *false negatives* (FN).

*Accuracy* reflects the overall fraction of correctly classified registrations (both accepted and rejected). *Sensitivity* (TPR), the fraction of correct registrations successfully accepted, captures the ability to identify successful registrations without incorrectly rejecting them. *Precision*, the fraction of accepted registrations that are actually correct, reflects decision quality when the system or operator chooses to accept, that is, how reliably accepted cases are truly correct. In contrast, *specificity* (TNR) measures the fraction of incorrect registrations that are correctly rejected, capturing the ability to identify and dismiss failed registrations and thus prevent false acceptances in safety-critical settings.

To systematically evaluate the Human+AI and Human+XAI conditions based on AI-derived outcomes, we constructed a balanced subset containing equal numbers of TP, TN, FP, and FN cases. While AI models can be evaluated on large datasets, user studies are limited by participant capacity; the balanced subset

**Table 1.** Descriptive Statistics (Mean±SD) for Performance Metrics Across Conditions

| Condition | Accuracy | Sensitivity (TPR) | Precision | Specificity (TNR) |
|---|---|---|---|---|
| Human-only | $0.547 \pm 0.158$ | $0.736 \pm 0.285$ | $0.531 \pm 0.248$ | $0.701 \pm 0.199$ |
| Human+AI | $0.677 \pm 0.125$ | $\mathbf{0.964 \pm 0.061}$ | $0.659 \pm 0.185$ | $0.848 \pm 0.079$ |
| Human+XAI | $\mathbf{0.678 \pm 0.124}$ | $0.877 \pm 0.239$ | $\mathbf{0.675 \pm 0.294}$ | $\mathbf{0.865 \pm 0.099}$ |

ensures a representative yet manageable set of cases for human evaluation. Although the balanced subset clarifies user responses to various AI errors, it does not reflect the actual error distribution. Thus, we computed approximate "real-world" metrics by weighting each category's fraction-correct by its prevalence in the entire test set ($TP$=22.6%, $TN$=53.4%, $FP$=18.8%, $FN$=5.1%).

**Subjective Measures** Subjective evaluations were collected to assess participants' perceived workload and their experience with AI assistance. Perceived workload was measured using the NASA-TLX questionnaire [5]. Each subscale was rated on a 21-point scale (0-100 in increments of 5), and the overall composite score was calculated by summing the six subscales and multiplying by 5/6 to obtain a score out of 100, following the Raw TLX approach. Additionally, participants rated trust in, usefulness of, and understanding of AI assistance in both Human+AI and Human+XAI conditions using a Likert-scale questionnaire (1: strongly disagree to 5: strongly agree). These subjective evaluations complement quantitative performance metrics by offering insights into participants' experiences.

To analyze differences across conditions, pairwise comparisons were conducted within subjects. Normality of paired differences was assessed with the Shapiro-Wilk test. For normally distributed differences ($p > 0.05$), paired t-tests were used; otherwise, Wilcoxon signed-rank tests were applied. To account for multiple comparisons, Holm-Bonferroni corrections were applied within each NASA-TLX metric family (three pairwise comparisons per metric). Statistical analyses were performed using Python's scipy library (v1.13.1).

## 3 Results

### 3.1 Performance Across Decision-Making Conditions

Table 1 summarizes the descriptive statistics of the performance metrics for each condition. Among the three conditions, Human+AI ($0.677 \pm 0.125$) and Human+XAI ($0.678 \pm 0.124$) achieved substantially higher *accuracy* than Human-only ($0.547 \pm 0.158$), representing a 24% relative improvement. For *sensitivity* (TPR), measuring the ability to correctly accept successful registrations, Human+AI achieved the highest value ($0.964 \pm 0.061$), followed by Human+XAI ($0.877 \pm 0.239$), both markedly exceeding Human-only ($0.736 \pm 0.285$).

In terms of *precision*, reflecting decision quality when accepting registrations, Human+XAI achieved the highest value ($0.675 \pm 0.294$), followed by Human+AI ($0.659 \pm 0.185$), both outperforming Human-only ($0.531 \pm 0.248$). For

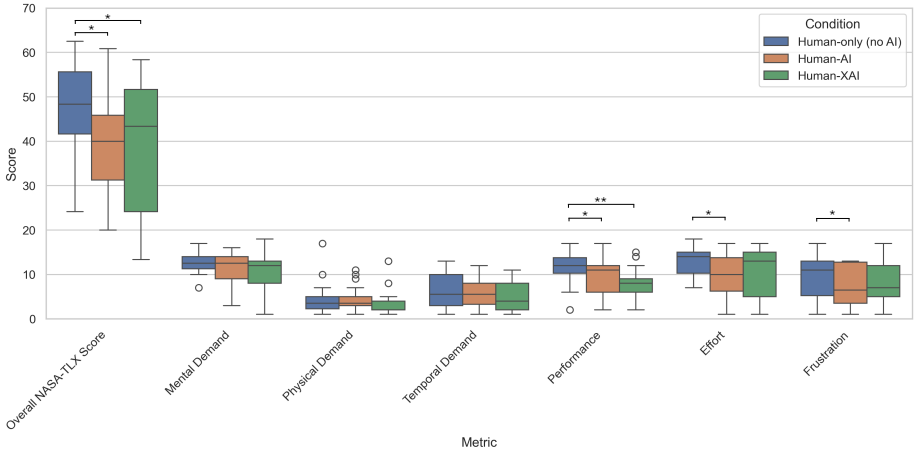


**Fig. 2.** Box plots of NASA-TLX scores across three conditions: (1) Human-only, (2) Human+AI, and (3) Human+XAI. Lower scores indicate lower perceived workload. Significance markers reflect Holm-Bonferroni corrected p-values within each metric. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

*specificity* (TNR), critical for avoiding false acceptances of failed registrations, Human+XAI achieved the highest mean ($0.865 \pm 0.099$), followed by Human+AI ($0.848 \pm 0.079$), compared to Human-only ($0.701 \pm 0.199$).

These results indicate that AI assistance improves human decision-making, particularly in correctly rejecting failed registrations (specificity) and accepting successful ones (sensitivity). The addition of explainability (Human+XAI) yielded the highest precision and specificity, suggesting explanations help users make more accurate accept/reject decisions.

### 3.2 Perceived Workload Across Human-AI Conditions

Figure 2 presents NASA-TLX scores measured across the three experimental conditions. Overall perceived workload, measured by the aggregate NASA-TLX score, decreased substantially in both AI-assisted conditions compared to Human-only. Specifically, workload decreased significantly in both Human+AI ($p = 0.0137$) and Human+XAI ($p = 0.0137$) when compared to Human-only, and no significant difference was observed between Human+AI and Human+XAI ($p = 0.3596$). Assessment of specific workload dimensions using pairwise testing yielded further insights. Effort decreased significantly in the Human+AI condition compared to Human-only ($p = 0.018$), and Frustration levels decreased significantly in Human+AI compared to Human-only ($p = 0.020$). Performance assessments, measured by participants' self-perceived effectiveness in task completion, were significantly higher in both Human+AI ($p = 0.042$) and Human+XAI

**Table 2.** Comparison of Subjective Measures between Human+AI and Human+XAI Conditions. Means (±SD) are shown, together with the statistical test, p-value, and effect size. Effect sizes are reported as absolute values; direction is indicated by mean differences (Cohen's $d$ for t-tests, rank-biserial $r$ for Wilcoxon tests). t = paired t-test; WSR = Wilcoxon signed-rank test; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

| Metric | Human+AI | Human+XAI | Test | p-value | Effect Size |
|---|---|---|---|---|---|
| Understanding | $2.83 \pm 1.01$ | $3.25 \pm 1.09$ | WSR | **0.030*** | $r = 0.86$ |
| Trust | $2.81 \pm 0.55$ | $3.06 \pm 1.06$ | t | 0.276 | $d = 0.27$ |
| Impact | $3.53 \pm 0.76$ | $3.64 \pm 0.97$ | t | 0.707 | $d = 0.09$ |
| Helpful | $3.83 \pm 0.49$ | $3.69 \pm 0.79$ | t | 0.472 | $d = 0.17$ |

($p = 0.001$) conditions relative to Human-only. No statistically significant differences were observed between Human+AI and Human+XAI conditions in Performance metrics ($p = 0.094$). Mental Demand, Physical Demand, and Temporal Demand showed no significant differences between conditions after correction.

### 3.3 Perceived Qualities of AI Assistance

Table 2 summarizes participants' subjective evaluations of AI assistance. No significant differences were identified for the metrics assessing perceived trust ($p = 0.276$), impact ($p = 0.707$), or helpfulness ($p = 0.472$). However, perceived understanding was significantly higher in the Human+XAI condition ($M = 3.25, SD = 1.09$) than the Human+AI condition ($M = 2.83, SD = 1.01$) with $p = 0.030$, indicating that explainability mechanisms enhanced participants' comprehension of AI predictions.

## 4 Discussion and Conclusion

We conducted a user study comparing three human decision-making paradigms for 2D/3D registration quality assurance: Human-only, Human+AI, and Human+XAI.

Both collaborative paradigms (Human+AI and Human+XAI) substantially outperformed Human-only in accuracy, sensitivity, precision, and specificity, demonstrating the value of AI support. Human+XAI achieved the highest precision and specificity, suggesting that explainability further supported participants in avoiding false acceptances. Clinically, these gains are meaningful: in surgical navigation, incorrectly accepting a misaligned registration carries greater consequence than conservatively rejecting a well-aligned one. We note that the prevalence-weighted performance metrics provide approximate real-world estimates, but assume participant behavior on the balanced subset would generalize to natural error distributions. Future work presenting cases with actual AI error rates would provide complementary validation of these findings.

Regarding subjective measures, participants reported significant reductions in perceived workload in both Human+AI and Human+XAI conditions compared to Human-only, suggesting that AI support effectively reduces cognitive effort in decision-making tasks. However, no significant difference in workload was identified between Human+AI and Human+XAI, indicating that explainability mechanisms did not add measurable cognitive burden. While participants in the Human+XAI condition provided higher ratings for understanding, no significant improvements over Human+AI were observed for perceived trust, impact, or helpfulness of AI assistance. These results may have been influenced by the balanced subset design. By ensuring equal numbers of correct and incorrect AI predictions, participants encountered AI errors more frequently than would occur in deployment. This likely contributed to the absence of a trust difference between Human+AI and Human+XAI. Under a realistic error distribution where AI is correct most of the time, trust outcomes could shift. Future work should validate these findings with the AI's actual error distributions.

Several limitations warrant discussion. First, our sample size (N=18), while adequate for detecting large effects, may lack power to identify smaller differences between Human+AI and Human+XAI conditions. The high variance observed in some metrics (e.g., sensitivity SD ranging from 0.061 to 0.285) suggests future studies with larger samples are needed to establish more precise effect estimates. Non-significant comparisons between collaborative conditions should therefore be interpreted as potentially power-limited rather than definitive null effects.

Several avenues remain open for future exploration. First, the interaction between human operators and XAI was relatively static in our study to enable controlled evaluation of how users respond to AI predictions and explanations. Future work could incorporate iterative or conversational interfaces, where operators can query the AI on uncertain areas for real-time feedback. The dissociation between improved understanding and unchanged objective performance suggests static Grad-CAM overlays may not surface the actionable information needed to revise decisions. Second, adopting gaze tracking [2] and other physiological or behavioral signals could offer deeper insights into how operators detect misalignments or interpret AI explanations. Third, as the work progresses toward clinical deployment, system-level metrics such as inference time and computational scalability should also be evaluated to ensure assessments fit within intraoperative workflow constraints.

Overall, our findings highlight the dual need for accurate models and well-designed explanations and interactions that truly support human judgment in safety-critical contexts. As surgical technologies continue to advance, it remains essential to pursue a human-centered perspective, ensuring they augment rather than supplant the expertise and accountability of the operating room team.

**Acknowledgments.** This study was supported in part by Johns Hopkins Internal Funds and by a Google Research Scholar Award.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References


1. Cho, S.M., Grupp, R.B., Gomez, C., Gupta, I., Armand, M., Osgood, G., Taylor, R.H., Unberath, M.: Visualization in 2d/3d registration matters for assuring technology-assisted image-guided surgery. IJCARS **18**(6), 1017–1024 (2023)
2. Cho, S.M., Taylor, R.H., Unberath, M.: Misjudging the machine: Gaze may forecast human-machine team performance in surgery. In: MICCAI. pp. 401–410. Springer (2024)
3. Cho, S.M., Zou, X., Fleig, L., Unberath, M.: Mini-review on human-centered assurance in robot-assisted orthopedics and neurosurgery. Frontiers in Robotics and AI **13**, 1755883
4. Fiorini, P., Goldberg, K.Y., Liu, Y., Taylor, R.H.: Concepts and trends in autonomy for robot-assisted surgery. Proceedings of the IEEE **110**(7), 993–1011 (2022)
5. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. In: Advances in psychology, vol. 52, pp. 139–183. Elsevier, North-Holland (1988)
6. Van de Kraats, E.B., Penney, G.P., Tomaževič, D., van Walsum, T., Niessen, W.J.: Standardized evaluation of 2d-3d registration. In: MICCAI. pp. 574–581. Springer (2004)
7. Markelj, P., Tomaževič, D., Likar, B., Pernuš, F.: A review of 3d/2d registration methods for image-guided interventions. Medical image analysis **16**(3), 642–661 (2012)
8. Mirota, D.J., Ishii, M., Hager, G.D.: Vision-based navigation in image-guided interventions. Annual review of biomedical engineering **13**, 297–319 (2011)
9. Rivero-Moreno, Y., Rodriguez, M., Losada-Muñoz, P., Redden, S., Lopez-Lezama, S., Vidal-Gallardo, A., Machado-Paled, D., Guilarte, J.C., Teran-Quintero, S.: Autonomous robotic surgery: Has the future arrived? Cureus **16**(1) (2024)
10. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
11. Unberath, M., Gao, C., Hu, Y., Judish, M., Taylor, R.H., Armand, M., Grupp, R.: The impact of machine learning on 2d/3d registration for image-guided interventions: A systematic review and perspective. Frontiers in Robotics and AI **8**, 716007 (2021)