Evaluation of the Clinical Efficacy and Trust in Artificial Intelligence-Assisted Embryo Ranking: A Survey-Based Prospective Study

Hyung Min Kim, Hyoeun Kang, Chaeyoon Lee, Jong Hyuk Park, Mi Kyung Chung, Miran Kim, Na Young Kim, Hye Jun Lee

Submitted to: Journal of Medical Internet Research
on: September 12, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Evaluation of the Clinical Efficacy and Trust in Artificial Intelligence-Assisted Embryo Ranking: A Survey-Based Prospective Study

Hyung Min Kim[1] Phd; Hyoeun Kang[1] MS; Chaeyoon Lee[1] BS; Jong Hyuk Park[2] PhD; Mi Kyung Chung[3] PhD; Miran Kim[4] MD, PhD; Na Young Kim[5] MD; Hye Jun Lee[6] MD

[1]AI Lab Kai Health Seoul KR
[2]IVF clinic Miraewaheemang Hospital Seoul KR
[3]IVF clinic Seoul Rachel Fertility Center Seoul KR
[4]Department of Obstetrics & Gynecology Ajou University School of Medicine Suwon KR
[5]IVF clinic HI fertility center Seoul KR
[6]Kai Health Chief executive officer Seoul KR

Corresponding Author:
Hye Jun Lee MD
Kai Health
Chief executive officer
217 Teheran-ro #306, Yeoksam-dong, Gangnam-gu
Seoul
KR

Abstract

Background: Current embryo assessment methods for in vitro fertilization (IVF) depend on subjective morphological assessments. Recently, artificial intelligence (AI) has emerged as a promising tool for embryo assessment; however, its clinical efficacy and trustworthiness remains unproven. Simulation studies may provide additional evidence, provided that they are meticulously designed to mitigate bias and variance.

Objective: The primary objective of this study was to evaluate the benefits of an AI model for predicting clinical pregnancy through well-designed simulations. The secondary objective was to identify the characteristics of and potential bias in the subgroups of embryologists with varying degrees of experience.

Methods: This simulation study involved a questionnaire based survey conducted on 61 embryologists with varying levels of experience from twelve IVF clinics. Inter- and intra-observer assessments and the accuracy of embryo selection from 360 day 5 embryos before and after AI guidance were analyzed for all embryologists and subgroups of senior and junior embryologists.

Results: With AI guidance, the inter-observer agreement increased from 0.355 to 0.527 and from 0.440 to 0.524 for junior and senior embryologists, respectively, thus reaching similar levels of agreement. The overall accuracies of the embryologists only, embryologists with AI guidance, and AI only were 37.7%, 50%, and 65.5%, respectively. Without AI, the average accuracy of the junior group was 33.516 (37.2%), while that of the senior group was 35.967 (40.0%). With AI's guidance, the junior group's accuracy improved to 46.581 (51.8%), reaching a level similar to that of the senior embryologists, 44.833 (49.8%). The junior embryologists had a higher level of trust in the AI score.

Conclusions: This study demonstrates the potential benefits of AI in selecting embryos with high chances of pregnancy, particularly for embryologists with less than or equal to 5 years of experience, possibly due to their trust in AI. Thus, using AI as an auxiliary tool in clinical practice has the potential to improve embryo assessment and increase the probability of a successful pregnancy.

(JMIR Preprints 12/09/2023:52637)
DOI: https://doi.org/10.2196/preprints.52637

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Evaluation of the Clinical Efficacy and Trust in Artificial Intelligence-Assisted Embryo Ranking: A Survey-Based Prospective Study

Hyung Min Kim[1], Hyoeun Kang[1], Chaeyoon Lee[1], Jong Hyuk Park[2], Mi Kyung Chung[3], Miran Kim[4], Na Young Kim[5], Hye Jun Lee[6]

[1] Kai Health, AI Lab, Seoul, South Korea
[2] Miraewaheemang Hospital, IVF clinic, Seoul, South Korea
[3] Seoul Rachel Fertility Center, IVF clinic, Seoul, South Korea
[4] Department of Obstetrics & Gynecology, Ajou University School of Medicine, Suwon, South Korea
[5] HI fertility center, IVF clinic, Seoul, South Korea
[6] Kai Health, Chief executive officer, Seoul, South Korea

Corresponding Author:
Hye Jun Lee
217 Teheran-ro #306, Yeoksam-dong, Gangnam-gu Seoul, 06142
South Korea
Phone: +82.10.5229.9697
Email: hyejunlee@gmail.com

Abstract

Background: Current embryo assessment methods for in vitro fertilization (IVF) depend on subjective morphological assessments. Recently, artificial intelligence (AI) has emerged as a promising tool for embryo assessment; however, its clinical efficacy and trustworthiness remains unproven. Simulation studies may provide additional evidence, provided that they are meticulously designed to mitigate bias and variance.
Objective: The primary objective of this study was to evaluate the benefits of an AI model for predicting clinical pregnancy through well-designed simulations. The secondary objective was to identify the characteristics of and potential bias in the subgroups of embryologists with varying degrees of experience.
Methods: This simulation study involved a questionnaire-based survey conducted on 61 embryologists with varying levels of experience from twelve IVF clinics. The survey was conducted via Google Forms (Google) in three phases: (1) Phase 1: initial assessment (December 23, 2022 to Jan 22, 2023), (2) Phase 2: validation assessment (March 6, 2023 to April 5, 2023), and (3) Phase 3: AI-guided assessment (March 6, 2023 to April 5, 2023). Inter- and intra-observer assessments and the accuracy of embryo selection from 360 day 5 embryos before and after AI guidance were analyzed for all embryologists and subgroups of senior and junior embryologists.
Results: With AI guidance, the inter-observer agreement increased from 0.355 to 0.527 and from 0.440 to 0.524 for junior and senior embryologists, respectively, thus reaching similar levels of agreement. The overall accuracies of the embryologists only, embryologists with AI guidance, and AI only were 37.7%, 50%, and 65.5%, respectively. Without AI, the average score (accuracy) of the junior group was 33.516 (37.2%), while that of the senior group was 35.967 (40.0%). With AI

guidance, the average score (accuracy) of the junior group increased to 46.581 (51.8%), reaching a level similar to that of the senior embryologists, 44.833 (49.8%). Junior embryologists had a higher level of trust in the AI score.

Conclusions: This study demonstrates the potential benefits of AI in selecting embryos with high chances of pregnancy, particularly for embryologists with less than or equal to 5 years of experience, possibly due to their trust in AI. Thus, using AI as an auxiliary tool in clinical practice has the potential to improve embryo assessment and increase the probability of a successful pregnancy.

Keywords: assisted reproductive technology; in vitro fertilization; artificial intelligence; intra-observer and inter-observer agreement

Introduction

Infertility affects one in six couples worldwide, making in vitro fertilization (IVF) a widely sought-after solution. However, the success rate of IVF remains relatively low, typically ranging from 20% to 30% [1,2]. Amidst the ongoing efforts to improve IVF outcomes, the paramount challenge lies in selecting the most viable embryo for transfer, as embryo quality is critical for a successful outcome.

Traditionally, embryologists rely on morphological assessment for selecting embryos [3–5], which involve the observation of embryos under a microscope and assigning grades based on criteria, such as blastocyst expansion stage, inner cell mass, and trophectoderm development. Some laboratories select euploid embryos through preimplantation genetic testing (PGT) [6]. Even after PGT, morphological assessment remains crucial to select the most viable one from multiple euploid embryos.

However, this reliance on morphological assessment raises concerns because of its inherent subjectivity and the substantial variability observed in both intra- and inter-observer assessments [7–11]. This variability underscores the pressing need for standardized methods of embryo evaluation across laboratories and the IVF industry.

Recent advancements have introduced artificial intelligence (AI) as a complementary tool for morphological evaluation of embryos. By leveraging deep learning techniques, AI-based systems can predict IVF outcomes by learning from extensive sets of embryo images, thus reducing human bias and potentially providing more objective and accurate results [12–17]. Several studies have reported significant advantages of AI-selected embryos in terms of pregnancy rates, compared to embryos chosen through traditional morphological assessment by embryologists [12,15,18].

Despite the notable progress in research on embryo selection using AI, its widespread adoption hinges on proving its clinical efficacy and securing the trust of clinicians. However, demonstrating the clinical efficacy of AI in a real-world setting poses challenges, as pregnancy outcomes can only be observed for selected embryos and not for those left unselected. Furthermore, the ultimate decision regarding which embryos to transfer rests squarely with clinicians, who may either embrace or question AI recommendations. In cases where clinicians opt not to trust AI, the resulting IVF outcomes may not accurately reflect the precision of the AI model’s predictions.

Previous attempts to address these critical questions have involved simulation studies that compared the accuracy of and pregnancy rates facilitated by AI-driven embryo selection and assessments by embryologists [19,20]. However, these studies predominantly relied on historical data from embryo grading records provided by embryologists from various laboratories. This retrospective approach inadvertently introduced substantial sources of variability. Intra- and inter-observer agreement among embryologists, which arise from the differing standards and criteria for embryo evaluation, played a significant role in shaping the results. Moreover, the lack of standardization of embryo evaluation across laboratories further complicates the interpretation of the findings.

In addition to these challenges, previous simulation studies have not examined the nuanced demographic profiles and attitudes of embryologists themselves. Understanding these factors can shed light on the broader dynamics at play and offer insights into how AI may be influenced by the unique characteristics and perspectives of the medical professionals involved.

In this study, we aimed to evaluate the clinical efficacy of AI in embryo selection by simulating a clinical setting in which embryologists ranked the embryos for transfer. We assessed the intra- and inter-observer agreement of the embryologists' evaluations to emphasize the need for alternative embryo selection methods. Moreover, we compared the accuracy of embryologists with and without AI assistance as well as AI-only selection to substantiate the clinical efficacy of AI. Additionally, we analyzed and compared the results for subgroups of embryologists with varying levels of experience to identify distinct practice patterns and levels of trust in AI. We believe that the results of this simulation study can serve as a foundational step for large-scale clinical investigations.

Methods

Study Design and Ethical Considerations

This study was a prospective cohort study in which an online questionnaire-based survey was conducted among embryologists with varying degrees of experience. The intra- and inter-observer agreement of the evaluation of 360 day-5 embryos by the embryologists and the utility of a self-developed AI tool were assessed as a reference for embryologists. The embryo images used in our study were collected from seven IVF clinics in accordance with the Declaration of Helsinki and the institutional review boards (IRBs) of Miraewaheemang Hospital (IRB No. 2022-RESEARCH-01), Good Moonhwa Hospital (IRB No. GMH-2022-01), the HI Fertility Center (IRB No. HIRB 2022-01), Seoul Rachel Fertility Center (IRB No. RTR-2022-01), Ajou University Hospital (IRB No. AJIRB-MED-MDB-21-716), Pusan National University Hospital (IRB No. 2204-003-113), and Seoul National University Bundang Hospital (IRB No. B-2208-772-104). Because this study used retrospective data collected through the IRBs of the aforementioned institutions, informed consent was waived, and the personal information contained in the data was de-identified.

Survey

The survey was conducted in three phases via Google Forms (Google): (1) Phase 1: initial assessment (December 23, 2022 to Jan 22, 2023), (2) Phase 2: validation assessment (March 6, 2023 to April 5, 2023), and (3) Phase 3: AI-guided assessment (March 6, 2023 to April 5, 2023). The participants received an email with the link to the Google Forms and attachment of original embryo images included in the questions in case they wanted an enlarged view. The submitted survey results were collected and used for analysis. The survey participants were compensated with US $20 for submission of the results.

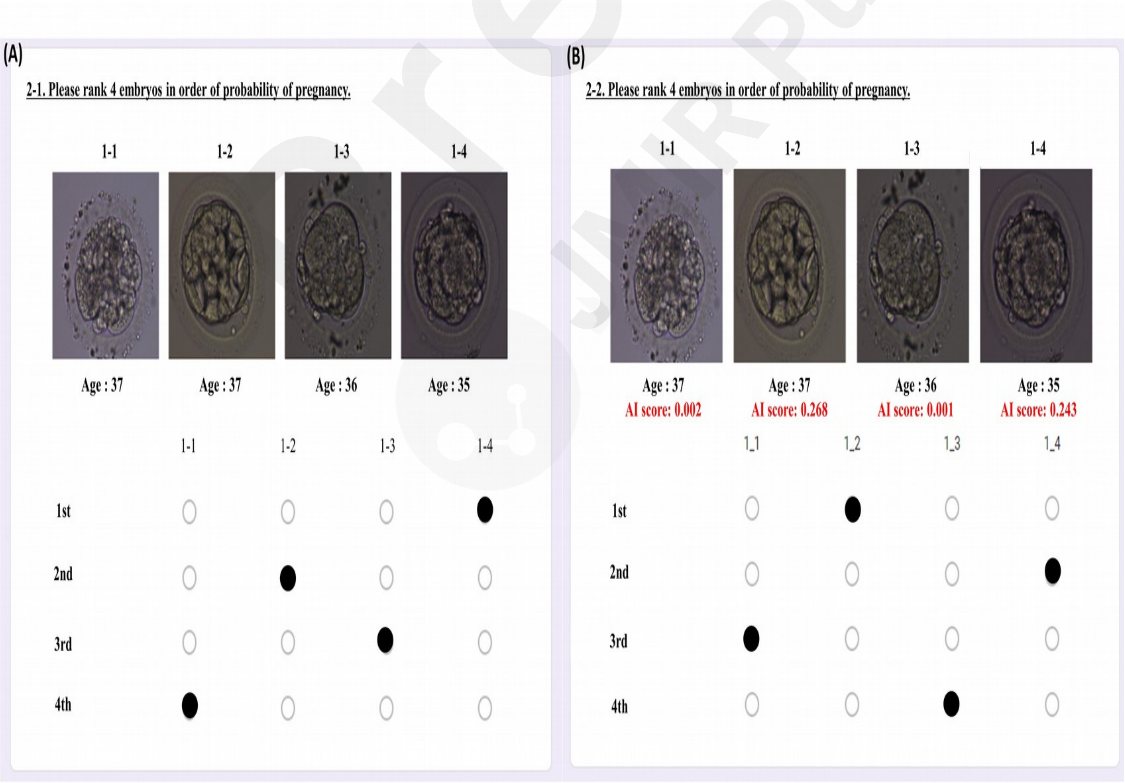
To measure intra- and inter-observer agreement, the initial assessment and the validation assessment were performed with one month apart and the questions in the initial assessment were identical to those in the validation assessment. Intra-observer agreement was analyzed based on the differences in responses between the initial and validation assessments (Phases 1 and 2), while inter-observer agreement was assessed based on the average of the responses. To assess the efficacy of AI guidance, AI-guided assessment (Phase 3) was performed right after the validation assessment.

The questionnaire was divided into two major sections: items designed to analyze the accuracy of embryo selection and demographic information (age, gender, highest educational level, and tenure). Each assessment consisted of 90 questions and each question consisted of images of three embryos that did not result in clinical pregnancy and one embryo that did. In the initial and validation assessments, the embryologists were asked to arrange the images in the order of the embryos with the highest likelihood of pregnancy. In the AI-guided assessment, AI scores were provided alongside the embryo images, and the embryologists were asked to reorder the embryos based on their perceived likelihood of pregnancy while considering the AI scores (Figure 1).

Out of 90 questions, 70 questions featured day-5 embryos at the same developmental stage and the remaining 20 questions used randomly selected day-5 embryos regardless of developmental stage to compare the effect of developmental stage of embryos in embryo selection. To minimize the effect of women's age, each question contained images of embryos from women of the same age group; under 37 and 37 and above.

After the survey, the embryologists were asked to express their opinions on the following aspects: (1) difficulty of the test and the reason for their opinion about the difficulty; (2) reason for the criteria that they considered when ranking the selected embryos (e.g., age, AI score, and image); (3) reason for changing the answer based on AI score; and (4) consideration when selecting embryos conventionally, as opposed to using a test approach.

Figure 1. Questions on embryo selection from the web survey. (A) Phase 2 of the survey: question on embryo selection without AI scores. (B) Phase 3 of the survey: question on embryo selection with AI scores.



(A)

2-1. Please rank 4 embryos in order of probability of pregnancy.

1-1	1-2	1-3	1-4
Age : 37	Age : 37	Age : 36	Age : 35

	1-1	1-2	1-3	1-4
1st	○	○	○	●
2nd	○	●	○	○
3rd	○	○	●	○
4th	●	○	○	○

(B)

2-2. Please rank 4 embryos in order of probability of pregnancy.

1-1	1-2	1-3	1-4
Age : 37 AI score: 0.002	Age : 37 AI score: 0.268	Age : 36 AI score: 0.001	Age : 35 AI score: 0.243

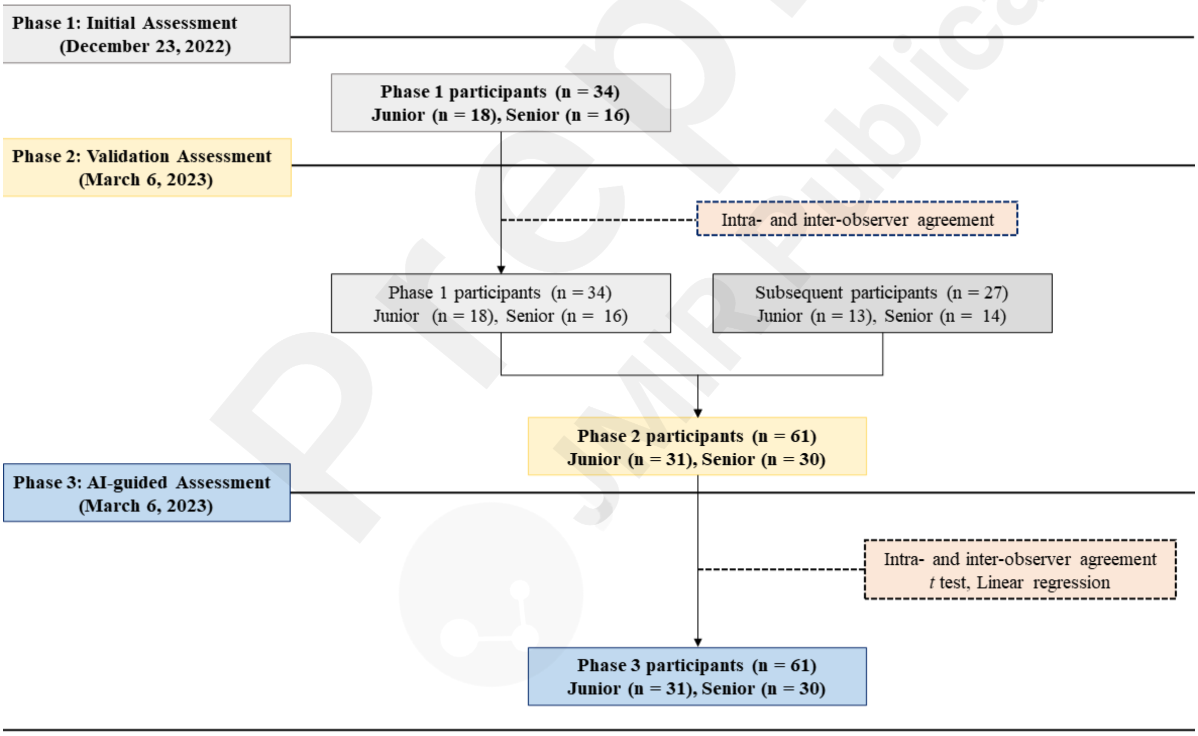
	1_1	1_2	1_3	1_4
1st	○	●	○	○
2nd	○	○	○	●
3rd	●	○	○	○
4th	○	○	●	○

Participants

We conducted initial assessment with 34 embryologists to build a baseline and included 27 more embryologists for validation and AI-guided assessment. The total 61 embryologists were recruited from 12 different IVF clinics and had clinical experience of one to over 30 years. All participants gave informed consent online before participating, and those who refused to give consent were excluded. Among them, 50 Korean embryologists were certified by the Korean Association for Clinical Embryologists and conducted an average of 40–150 cases per month. Of the remaining embryologists, nine were from Malaysia and two were from the United States of America. These participants were divided into two groups: (1) junior group consisting of embryologists with embryo grading experience ≤5 years and (2) senior group consisting of embryologists with embryo grading experience ¿5 years.

All 61 participants successfully completed the validation and AI-guided assessments without any loss to follow-up. To analyze the results of the validation and AI-guided assessments conducted before and after referring to the AI scores, 29 additional participants were recruited. Statistical analyses were conducted on all participants together and between groups (Figure 2), respectively. From the junior group, 18, 31, and 31 embryologists participated in the initial, validation, and AI-guided assessments, respectively, and from the senior group 16, 30, and 30 embryologists participated in the initial, validation, and AI-guided assessments, respectively.

Figure 2. Study design and participant flow diagram according to the survey phase


Phase 1: Initial Assessment (December 23, 2022)
Phase 1 participants (n = 34) Junior (n = 18), Senior (n = 16)
Phase 2: Validation Assessment (March 6, 2023)
Intra- and inter-observer agreement
Phase 1 participants (n = 34) Junior (n = 18), Senior (n = 16)
Subsequent participants (n = 27) Junior (n = 13), Senior (n = 14)
Phase 2 participants (n = 61) Junior (n = 31), Senior (n = 30)
Phase 3: AI-guided Assessment (March 6, 2023)
Intra- and inter-observer agreement t test, Linear regression
Phase 3 participants (n = 61) Junior (n = 31), Senior (n = 30)

Evaluation of the Impact of AI Guidance on Embryo Selection

We measured the accuracy of embryo selection over four cycles to determine how much it increased when the embryologists were guided by AI. The first cycle refers to the case in which the embryo selected as rank 1 resulted in clinical pregnancy, and the second cycle refers to the case in which the embryo selected as rank 1 or 2 resulted in clinical pregnancy. Likewise, third cycle means that the embryo selected as 1, 2 or 3 resulted in clinical pregnancy, and fourth cycle means 100% because there are 4 examples per question.

Evaluation of Trust in AI

The model employed to infer the AI scores per image was trained on 2555 day-5 embryo images collected from seven Korean IVF clinics. A total of 2,555 images were divided into two sets: a training dataset consisting of 2,043 images (80%) and a model performance test dataset containing 512 images (20%). We then employed a 3-fold cross-validation approach to further divide the 2,043 training images into three separate folds. Each fold was used for both training and validation of the model, and performance was assessed using a fixed model performance test dataset. When trained using the ResNet50 architecture, the performance resulted in an area under the receiver operating characteristic of 0.716 and accuracy of 0.663 [21]. 360 embryo images used in this survey were extracted from 512 images in the dataset that was not used to train the AI model. For our questions in the clinical study, the accuracy was 65.5%, which is similar to the accuracy of the model test set. In assessing the accuracy of the AI model's predictions, we employed a method centered on the alignment of AI scores with the actual clinical outcomes. For each set of embryo images, encompassing four images per case, our criterion for determining accuracy was the extent to which the AI's highest score matched with the embryo that led to a successful pregnancy.

To determine the extent to which embryologists relied on AI scores to modify their responses, we defined the AI trust level as follows:

$$\text{AI trust level} = \text{Number of questions} \in \text{which embryologists revised their response} \text{¿} \text{higher AI score} \frac{}{\text{Number of questions} \in \text{which}}$$

The above formula resulted in a trust level of 0–1 for each embryologist. A level of 0 indicates that the embryologist performed all modifications to their responses on a subjective basis with no reference to the AI, whereas a level of 1 indicates that the embryologist performed all modifications so that their ranking was as consistent with the AI score as possible.

Statistical Analysis

Statistical analyses were performed to assess the consistency of embryo scoring between the embryologists. Cohen’s kappa coefficient was used to evaluate the intra-observer agreement between the scores given by the same embryologist at two different time points [22], whereas Fleiss kappa coefficient was used for inter-observer agreement between the scores given by different embryologists [23]. The kappa coefficient was subsequently construed as excellent (≥0.80), good (0.60–0.79), moderate (0.40–0.59), poor (0.20–0.39), or very poor (<0.20) levels of intra- and inter-observer agreement [24]. In this study, we used a t test and linear regression to compare the selection accuracy and AI trust level between junior and senior groups. IBM SPSS Statistics for Windows (Version 29.0, IBM Corporation, Armonk, NY, USA) was employed to quantify intra- and inter-observer agreement, while the Python Programming Language (Version 3.8.0, Wilmington, DE, USA) was used to conduct t tests and linear regression analyses. When performing the t test, as our data consisted of more than 30 samples per group, we assumed a normal distribution based on the central limit theorem (CLT).

Results

Demographics

The demographics of the two groups (Table 1) show that the junior group comprised 80.6% females and 19.4% males. The most common age subgroup of the junior group was the 30s (51.6%), followed by the 20s (45.2%) and 40s (3.2%). The senior group, on the other hand, had a proportion of 70.0% females and 30.0% males, with the largest proportion of embryologists (33.3%) being in their 40s and 50s, followed by the 30s (30.0%) and 20s (3.3%). Over half of the junior and senior groups (58.1% and 50.0%, respectively) had a master's degree. The junior group comprised 42.9% university graduates and no doctoral degree holders. The senior group had a high percentage of doctoral degree holders (33.3%), followed by bachelor's degrees (16.7%). In relation to embryo assessment expertise, the junior group had an average of 1.6 years of experience, while the senior group had an average of 13.2 years of experience.

Table 1. Demographic characteristics of the junior and senior groups.

		Junior (n=31)	Senior (n=30)
Gender	Male	6 (19.4%)	9 (30.0%)
	Female	25 (80.6%)	21 (70.0%)
Age group	20-29	14 (45.2%)	1 (3.3%)
	30-39	16 (51.6%)	9 (30.0%)
	40-49	1 (3.2%)	10 (33.3%)
	≥50	0 (0.0%)	10 (33.3%)
Highest level of education	Bachelor's degree	13 (42.9%)	5 (16.7%)
	Master's degree	18 (58.1%)	15 (50.0%)
	Doctoral degree	0 (0.0%)	10 (33.3%)
Experience in embryo selection (years)		1.6 ± 1.9	13.2 ± 7.4

Intra- and Inter-observer Agreements

The evaluation of intra-observer agreement revealed a Cohen kappa score of 0.662 between the initial and validation assessments, indicating good (0.60–0.79) concordance between the response of one embryologist at two separate time points. The coefficients for the junior and senior groups were 0.659 and 0.664, respectively, indicating that the less experienced group was able to provide consistent responses at a similar level as the experienced group (Table 2). Additionally, the correlation coefficients between the validation and AI-guided assessments were 0.735 for the overall population and 0.698 and 0.773 for the junior and senior groups, respectively. This indicates that all participants showed improved consistency after the AI guidance.

Table 2. Results of the evaluation of the intra- and inter-observer agreements.

	Intra-observer agreement: Cohen's kappa coefficient (95% confidence interval)		Inter-observer agreement: Fleiss' kappa coefficient (95% confidence interval)	
	Without AI (phase 1 vs. phase 2)	With AI vs. Without AI (phase 2 vs. phase 3)	Without AI (phase 2)	With AI (phase 3)
Overall	0.662 (0.631–0.692)	0.735 (0.670–0.770)	0.392 (0.389–0.395)	0.521 (0.518–0.524)
Junior	0.659 (0.603–0.714)	0.698 (0.653–0.744)	0.355 (0.349–0.360)	0.527 (0.521–0.532)
Senior	0.664 (0.604–0.710)	0.773 (0.722–0.825)	0.440 (0.434–0.445)	0.524 (0.518–0.529)

Phase 1: initial assessment, phase 2: validation assessment, phase 3: AI-guided assessment.

Next, we measured the Fleiss kappa coefficient for inter-observer agreement between multiple embryologists and found that validation assessment showed a poor coefficient (0.20–0.39) of 0.392 for the total population. Inter-observer agreement within the junior group was indicated by a poor coefficient of 0.355, whereas within the senior group, the agreement was indicated by a moderate coefficient of 0.440 (Table 2). After referring to the AI, the coefficient was moderate at 0.521 for the entire population, and the junior and senior groups showed improved concordance of 0.527 and 0.523, respectively. This result suggests that the junior group could make judgements with consistency that was similar to those of the senior group after the AI's guidance.

Impact of AI Guidance on Embryo Selection

In the first cycle, the accuracy of the embryologist was 37.7% and that of the AI model was 65.5% (Table 3). When the embryologist was guided by the AI score, the accuracy rate increased to 50%. The AI model outperformed embryologists in selecting an embryo that led to pregnancy by 27.8%, and embryologists with AI guidance outperformed embryologists without AI guidance by 12.3%. The difference in accuracy between the embryologists and AI model was 22.2% in the second cycle and 14.4% in the third cycle, and the performance gap between the embryologists and AI was reduced to 12.2% in the second cycle and 7.8% in the third cycle.

Table 3. Comparison of cumulative accuracies of embryologists, embryologists with AI guidance, and AI for the prediction of clinical pregnancy.

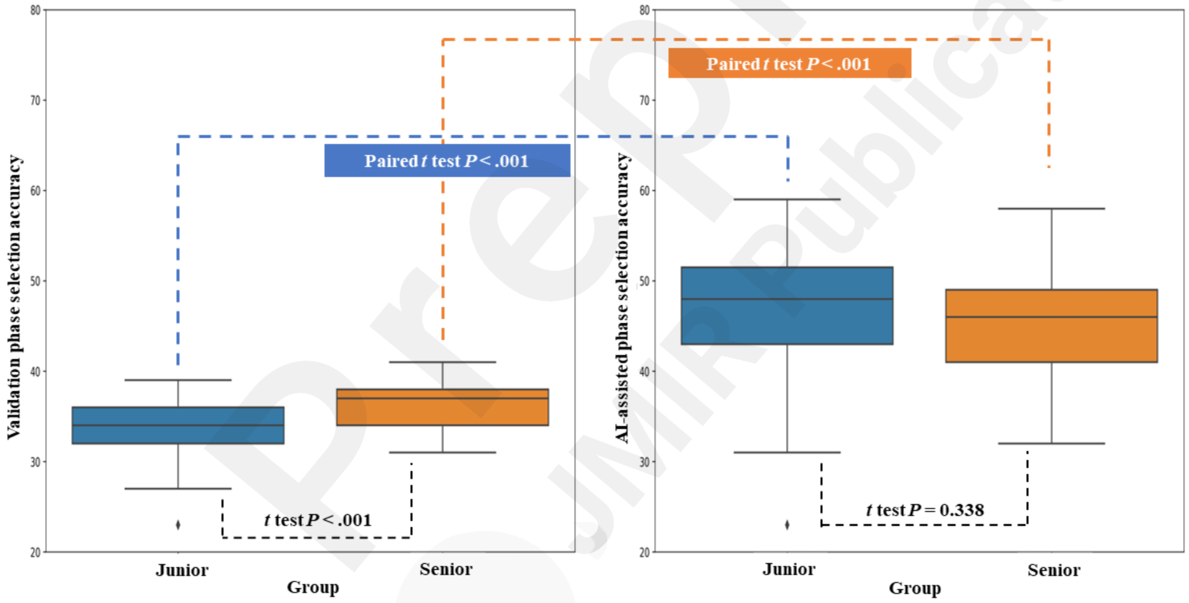
	Cumulative accuracy, n (%)			
	First cycle	Second cycle	Third cycle	Fourth cycle
Embryologists'	34 (37.7%)	57 (63.3%)	73 (81.1%)	90 (100%)

selection				
Embryologists' selection with AI guidance	45 (50%)	66 (73.3%)	79 (87.7%)	90 (100%)
AI selection	59 (65.5%)	77 (85.5%)	86 (95.5%)	90 (100%)

Relationship Between AI Trust Level and Embryo Selection Accuracy

The relationship between AI trust level and embryo selection accuracy was determined through statistical analysis of the accuracy of embryo selection by the junior and senior groups. The analysis revealed that in the validation assessment, the responses of the junior group differed significantly from those of the senior groups, with a *P* value of <.001 in the *t* test (Figure 3). The average accuracy of embryo selection by the junior group was 33.516 (SD, 3.688), while that of the senior group was 35.967 (SD, 2.580), indicating that embryologists with over 5 years of experience had significantly higher embryo selection ability.

Figure 3. Within- and between-group *t* test results



When comparing the two groups in the AI-guided assessment, the mean score of the junior group was 46.581 (SD, 7.967), and the mean score of the senior group was 44.833 (SD, 6.772), with *P*=.338, showing no significant difference between the two groups. In addition, a paired *t* test was performed between the validation and AI-guided assessments to determine whether there was a significant difference in selection accuracy before and after the groups referred to the AI score. The *P* value was less than .001 for both groups, confirming that the score increased after referring to AI.

Before checking the relationship between the AI trust level and embryo selection accuracy, we tested whether there was a difference in AI trust levels between the two groups. The AI trust level of the junior group was 0.581 (SD, 0.244), whereas that of the senior group was 0.443 (SD, 0.278). The *P* value of the *t* test was 0.047, which confirmed that the confidence of the junior group was significantly higher than that of the senior group (Figure 4). Subsequently, we performed a regression analysis with AI trust level as the independent variable and embryo selection score as the dependent variable. We found that the scores of both groups increased with the increase in the trust level. In addition, the regression coefficient of the junior group was 29.209, compared to 22.870 for the senior group; therefore, the slope of the embryo selection score with trust level increased sharply (Table 4). We further examined the distribution of the denominator in the AI trust level calculation by group. The distribution spanned from the smallest (29) to the largest number of different questions (61). We found that the junior group had 4 cases (12.9%) in which the top-ranked embryo was in the category in which 40 or fewer questions differed from the AI, while the senior group had 7 cases (23.3%). This suggests that the senior embryologists gave a higher number of answers that were similar to those of the AI. (Multimedia Appendix 1).

Figure 4. Results of the groupwise analyses of AI trust level through *t* test and regression analysis.

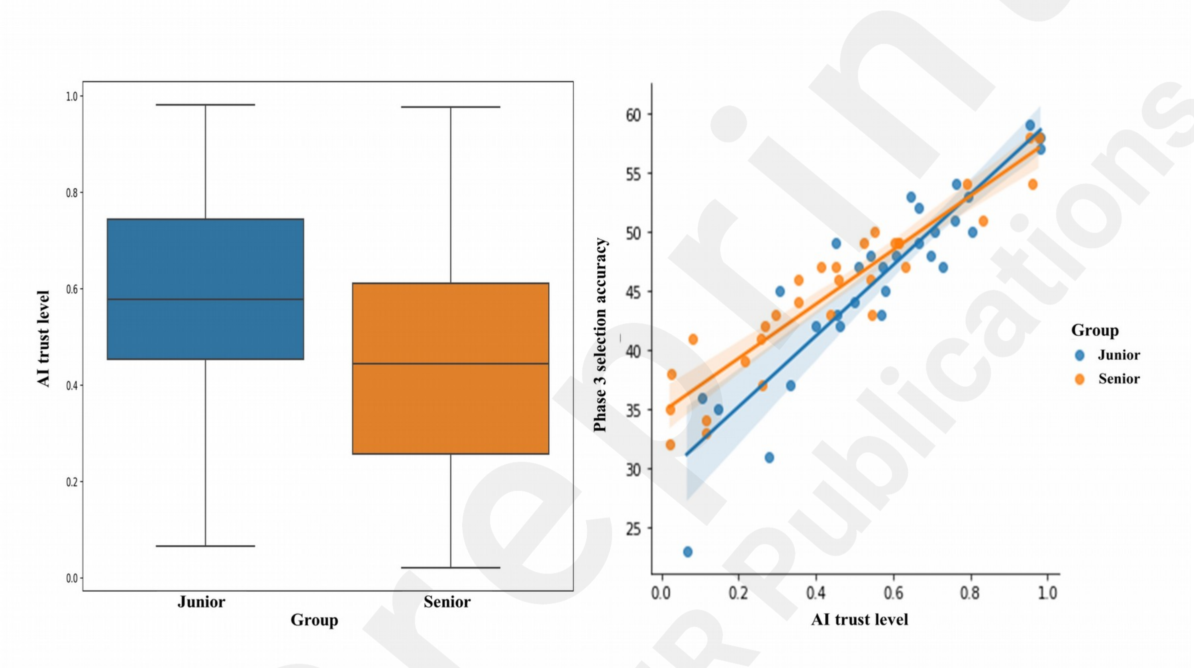


Table 4. Results of the groupwise regression analysis.

	Coefficient	Standard error	*P* value	95% confidence interval
Junior embryologist	29.209	1.514	<.001	24.996–34.821
Senior embryologist	22.870	1.605	<.001	19.582–26.158

Discussion

Principal Results

The primary objective of this study was to demonstrate the potential clinical benefits of integrating AI into the embryo-ranking process, ultimately increasing the likelihood of achieving a clinical pregnancy. The area under the receiver operating characteristic curve and accuracy of the model used in this study were 0.716 and 0.663, respectively. This was comparable to previous AI models developed upon 2D static images, with an accuracy of 0.64 [12] and an AUROC ranging from 0.6 to

0.7 [25]. In this study, we demonstrated that the AI model’s accuracy of 65.5% outperformed that of the embryologists, with 37.7%. Previous studies used the historical data of embryologists’ morphological grading to simulate embryo selection and found that the embryologists’ accuracy was between 0.47 and 0.65. Our study recruited a large number of embryologists rather than utilizing the historical data, which might have resulted the difference in the accuracy. Our investigation reveals several critical findings that shed light on the role of AI in this context.

First, our study revealed that intra- and inter-observer agreements among embryologists in ranking embryos improved with the assistance of AI. It is noted that the embryologists, particularly juniors, exhibited relatively low inter-observer agreement, but this was mitigated by the AI's guidance, effectively leveling the performance between juniors and seniors. Regarding intra-observer agreement, while the Cohen kappa score of 0.662 is statistically considered good (0.60–0.79), its clinical implications may differ, as evidenced by embryologists changing their responses to identical questions in 25% of questions over one month.

This variability underscored the need for a comparative analysis to assess the accuracy of embryo selection before and after the introduction of AI guidance. We utilized an AI model that demonstrated industrially standard performance and achieved an area under the receiver operating characteristic curve of 0.716 [21]. This performance metric aligns with previous studies that employed 2D images of day-5 embryos [12,25]. Furthermore, the positive correlation between AI scores and traditional manual grading by embryologists reinforced trust in the AI model (Multimedia Appendix 2).

To further substantiate the clinical benefits of AI, we conducted a blinded test in which embryologists ranked embryos without knowing their future outcomes. The findings were noteworthy as they indicated that the highest accuracy of selecting the most viable embryos was achieved by AI models followed by embryologists with AI guidance and embryologists without AI guidance. This observation suggests that our AI model has the potential to assist embryologists in the selection of the most viable embryos, thereby increasing the probability of successful pregnancies per cycle, while potentially reducing the time to conception.

This study was designed to closely simulate a clinical setting. Unlike previous simulation studies [19,25], we leveraged the embryologists' actual rankings rather than the rankings derived from their historical grading records. Although the morphological evaluation methodology is well established, the criteria for grading vary, resulting in limited intra- and inter-observer agreement. Previous studies have used manual grades as numeric scores mapped from alphanumeric historical grades [25] or employed random and Gardner-based scores as proxies for embryological accuracy [19]. These approaches face challenges in translating grades into ranks owing to the nonlinear nature of the embryo grading system.

Furthermore, we compared three distinct scenarios: embryologist-only, embryologists with AI guidance, and AI-only rankings. Previous studies have predominantly focused on comparing the embryologists’ independent assessments with AI-only evaluations [12,13]. However, it is imperative to include a scenario in which embryologists are guided by AI, as this closely mirrors the most likely clinical scenario, in which AI aids, rather than replaces, human judgement due to liability concerns.

In addition, we controlled for blastocyst developmental stages in 70 of 90 questions and maintained consistent age groups across all questions, allowing us to compare the outcomes between stage-controlled and random-stage questions. This approach closely mirrored the clinical context of embryo ranking. We observed the following accuracy rates: 32% and 61% without AI, 44% and 71% with AI guidance, and 60% and 85% for AI-only rankings in 70 blastocyst stage-adjusted questions and 20 random-stage questions, respectively. Notably, the accuracy of the questions involving randomly selected embryos with different blastocyst stages exceeded that of the stage-adjusted questions. This observation suggests that AI may offer the most substantial benefits in scenarios where embryologists frequently encounter assessment challenges. Our research design, which focused on embryos at similar stages, proved to be the most suitable for evaluating the clinical efficacy of the AI model.

A comprehensive questionnaire was also administered to gain deeper insight. The survey revealed significantly higher levels of trust in AI among junior embryologists than among their senior counterparts. Although junior embryologists initially exhibited lower accuracy rates than their senior peers before AI guidance, their performance improved and converged with those of their seniors after the AI intervention. An intriguing trend emerged from the regression analysis of confidence: for every one-unit increase in confidence, the junior group demonstrated a more substantial increase of 29.2 points, compared to the senior group’s increase of 22.87 points (Table 4). Interestingly, the current level of trust in AI appeared relatively modest, with 61.7% of the surveyed embryologists indicating that their ranking considerations included embryo morphology, age, and AI score. In contrast, 28.3% of the surveyed embryologists prioritized embryo morphology, AI scores, and age in their ranking considerations. This underscores the need for further research to establish clinical efficacy and foster trust among embryologists, particularly their senior counterparts.

Limitations

The AI model that we developed is highly effective in analyzing 2D static images. In the practical context of embryo selection, embryologists can assess embryos from multiple perspectives under a microscope. However, our experiment necessitated judgements based on single images captured from a single viewpoint. Additionally, our dataset comprised images captured by embryologists before embryo transfer; consequently, we lacked comprehensive kinetic information throughout the entire developmental process. Given the above limitations, we considered studies that covered complete embryonic development, such as time-lapse video analyses [26–30]. However, this method presents a set of challenges. Time-lapse equipment is expensive and requires embryologists to visually monitor the entire process, which requires considerable time and effort. Interestingly, prior research has suggested that using the final image taken on day 5 yields a predictive performance for pregnancy outcomes similar to that achieved with time-lapse images capturing the entire developmental process [31]. This insight led us to make a strategic decision to leverage our expertise in 2D image analysis to design a cost-effective and time-efficient experimental setup.

All the embryologists surveyed in this study reported varying levels of difficulty in the embryo ranking task, with the majority describing it as slightly difficult (52%) or moderately difficult (36%). Their perceived reasons for this difficulty include factors such as image quality and fixed focus. The embryo images used in this study were collected from various IVF clinics by introducing variations in magnification and color. This variability may have contributed to the less precise responses, as embryologists selected embryos under conditions that differed from standard practices. Therefore, for more accurate comparisons between AI and embryologists, future experiments should be conducted by collecting images of uniform size, magnification, and color within a single institution.

Conclusions

To date, there is a lack of practical research on the extent to which AI can assist researchers in embryo selection. In this study, we demonstrated that AI is crucial for successfully selecting embryos that provide high chances of pregnancy. This effect was particularly pronounced among embryologists with less than five years of experience who had more trust in AI scores. Thus, this study suggests that using AI as an auxiliary tool in clinical practice has the potential to enhance embryo assessment and increase the probability of a successful pregnancy.

Acknowledgements

We thank the 61 embryologists who participated in the survey.

Funding Statement

This study was supported by the Starting Growth Technological R&D Program (TIPS Program, (No. S3318926)) funded by the Ministry of SMEs and Startups (MSS, Korea) in 2022, as well as by the Korea Medical Device Development Fund grant funded by the Korean government (Ministry of Science and ICT, Ministry of Trade, Industry and Energy, Ministry of Health & Welfare, and Ministry of Food and Drug Safety) (RS-2023-00254175).

Data Availability

This study used the datasets from The Open AI Dataset Project (AI Hub, South Korea). All data can be accessed through ‘AI-Hub’ (www.aihub.or.kr).

Authors Contributions

Conceptualization: H.M.K. and H.J.L. Methodology: H.M.K. Software: H.M.K. and C.L. Validation: J.H.P. M.K.C., N.Y.K., and M.K. Formal Analysis: H.M.K. Acquisition of data: J.H.P., M.K.C., M.Y.K., and M.K. Writing—Original Draft Preparation: H.M.K. and H.K. Writing—Review and Editing: All authors. Visualization: H.M.K. Supervision: H.J.L. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

No potential conflict of interest relevant to this article is reported.
H.M.K, H.K, C.L, M.K and N.Y.K have no conflicts of interest to disclose. J.H.P and M.K .C report consulting fees from Kai Health. H.J.L reports stock options from Kai Health.

Abbreviations

AI: artificial intelligence
PGT: preimplantation genetic testing
IRB: institutional review board
IVF: in vitro fertilization
SD: standard deviation

Multimedia Appendix 1

Distribution of questions in which top-ranked embryos differ between embryologists and AI by group.

Multimedia Appendix 2

Correlation between AI scores and traditional manual grading.

References

1. Wang J, Sauer M V. In vitro fertilization (IVF): A review of 3 decades of

clinical innovation and technological advancement. Ther Clin Risk Manag 2006;2(4):355–364. PMID:18360648

2. Chow DJX, Wijesinghe P, Dholakia K, Dunning KR. Does artificial intelligence have a role in the IVF clinic? Reprod Fertil Bioscientifica Ltd; 2021 Sep 15;2(3):C29–C34. PMID:35118395
3. Racowsky C, Vernon M, Mayer J, Ball GD, Behr B, Pomeroy KO, Wininger D, Gibbons W, Conaghan J, Stern JE. Standardization of grading embryo morphology. J Assist Reprod Genet 2010;27(8):437–439. PMID:20532975
4. Balaban B, Brison D, Calderón G, Catt J, Conaghan J, Cowan L, Ebner T, Gardner D, Hardarson T, Lundin K, Cristina Magli M, Mortimer D, Mortimer S, Munné S, Royere D, Scott L, Smitz J, Thornhill A, Van Blerkom J, Van Den Abbeel E. The Istanbul consensus workshop on embryo assessment: Proceedings of an expert meeting. Hum Reprod 2011;26(6):1270–1283. PMID:21502182
5. Hill MJ, Richter KS, Heitmann RJ, Graham JR, Tucker MJ, Decherney AH, Browne PE, Levens ED. Trophectoderm grade predicts outcomes of single-blastocyst transfers. Fertil Steril Elsevier; 2013 Apr 1;99(5):1283-1289.e1. doi: 10.1016/J.FERTNSTERT.2012.12.003
6. Cornelisse S, Zagers M, Kostova E, Fleischer K, van Wely M, Mastenbroek S. Preimplantation genetic testing for aneuploidies (abnormal number of chromosomes) in in vitro fertilisation. Cochrane Database Syst Rev John Wiley and Sons, Inc. and the Cochrane Library; 2020 Sep 8;2020(9). PMID:32898291
7. Storr A, Venetis CA, Cooke S, Kilani S, Ledger W. Inter-observer and intra-observer agreement between embryologists during selection of a single Day 5 embryo for transfer: A multicenter study. Hum Reprod 2017;32(2):307–314. PMID:28031323
8. Bormann CL, Thirumalaraju P, Kanakasabapathy MK, Kandula H, Souter I, Dimitriadis I, Gupta R, Pooniwala R, Shafiee H. Consistency and objectivity of automated embryo assessments using deep neural networks. Fertil Steril 2020;113(4):781-787.e1. PMID:32228880
9. Adolfsson E, Andershed AN. Morphology vs morphokinetics: A retrospective comparison of interobserver and intra-observer agreement between embryologists on blastocysts with known implantation outcome. J Bras Reprod Assist 2018;22(3):228–237. PMID:29912521
10. Paternot G, Devroe J, Debrock S, D’Hooghe TM, Spiessens C. Intra- and inter-observer analysis in the morphological assessment of early-stage embryos. Reprod Biol Endocrinol BioMed Central; 2009 Sep 29;7(1):105. PMID:19788739
11. Sundvall L, Ingerslev HJ, Breth Knudsen U, Kirkegaard K. Inter- and intra-observer variability of time-lapse annotations. Hum Reprod Oxford Academic; 2013 Dec 1;28(12):3215–3221. PMID:24070998
12. Ver Milyea M, Hall JMM, Diakiw SM, Johnston A, Nguyen T, Perugini D, Miller A, Picou A, Murphy AP, Perugini M. Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. Hum Reprod 2020;35(4):770–784. PMID:32240301
13. Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, Sigaras A, Lavery S, Cooper LAD, Hickman C, Meseguer M, Rosenwaks Z,

Elemento O, Zaninovic N, Hajirasouliha I. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. npj Digit Med Springer US; 2019;2(1):1–9. doi: 10.1038/s41746-019-0096-y

14. Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. Reprod Biomed Online Elsevier Ltd; 2020 Oct 1;41(4):585–593. PMID:32843306

15. Bormann CL, Kanakasabapathy MK, Thirumalaraju P, Gupta R, Pooniwala R, Kandula H, Hariton E, Souter I, Dimitriadis I, Ramirez LB, Curchoe CL, Swain J, Boehnlein LM, Shafiee H. Performance of a deep learning based neural network in the selection of human blastocysts for implantation. Elife 2020;9:1–14. PMID:32930094

16. Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Hum Reprod 2019;34(6):1011–1018. PMID:31111884

17. Zaninovic N, Rosenwaks Z. Artificial intelligence in human in vitro fertilization and embryology. Fertil Steril 2020;114(5):914–920. PMID:33160513

18. Friedenthal J, Hernandez-Nieto C, Roth RM, Slifkin R, Gounko D, Lee JA, Nazem T, Briton-Jones C, Copperman A. Clinical implementation of algorithm-based embryo selection is associated with improved pregnancy outcomes in single vitrified warmed euploid embryo transfers. J Assist Reprod Genet Springer; 2021 Jul 1;38(7):1647–1653. PMID:33932196

19. Diakiw SM, Hall JMM, VerMilyea M, Lim AYX, Quangkananurug W, Chanchamroen S, Bankowski B, Stones R, Storr A, Miller A, Adaniya G, van Tol RA, Hanson R, Aizpurua J, Giardini L, Johnston A, Van Nguyen T, Dakka MA, Perugini D, Perugini M. An artificial intelligence model correlated with morphological and genetic features of blastocyst quality improves ranking of viable embryos. Reprod Biomed Online Elsevier Ltd; 2022;45(6):1105–1117. PMID:36117079

20. Bori L, Meseguer F, Valera MA, Galan A, Remohi J, Meseguer M. The higher the score, the better the clinical outcome: retrospective evaluation of automatic embryo grading as a support tool for embryo selection in IVF laboratories. Hum Reprod Hum Reprod; 2022 Jun 1;37(6):1148–1160. PMID:35435210

21. Kim HM, Ko T, Kang H, Choi S, Park JH, Chung MK, Kim M, Kim NY, Lee HJ. Improved Prediction of Clinical Pregnancy Using Artificial Intelligence with Enhanced Inner Cell Mass and Trophectoderm Images. 2023 Aug 8; doi: 10.21203/RS.3.RS-3204889/V1

22. Rau G, Shih YS. Evaluation of Cohen’s kappa and other measures of inter-rater agreement for genre analysis and other nominal data. J English Acad Purp Elsevier Ltd; 2021;53(June):101026. doi: 10.1016/j.jeap.2021.101026

23. Rücker G, Schimek-Jasch T, Nestle U. Measuring inter-observer agreement in contour delineation of medical imaging in a dummy run using fleiss’ kappa. Methods Inf Med Schattauer GmbH; 2012;51(6):489–494. PMID:23160666

24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics JSTOR; 1977 Mar 1;33(1):159–174.

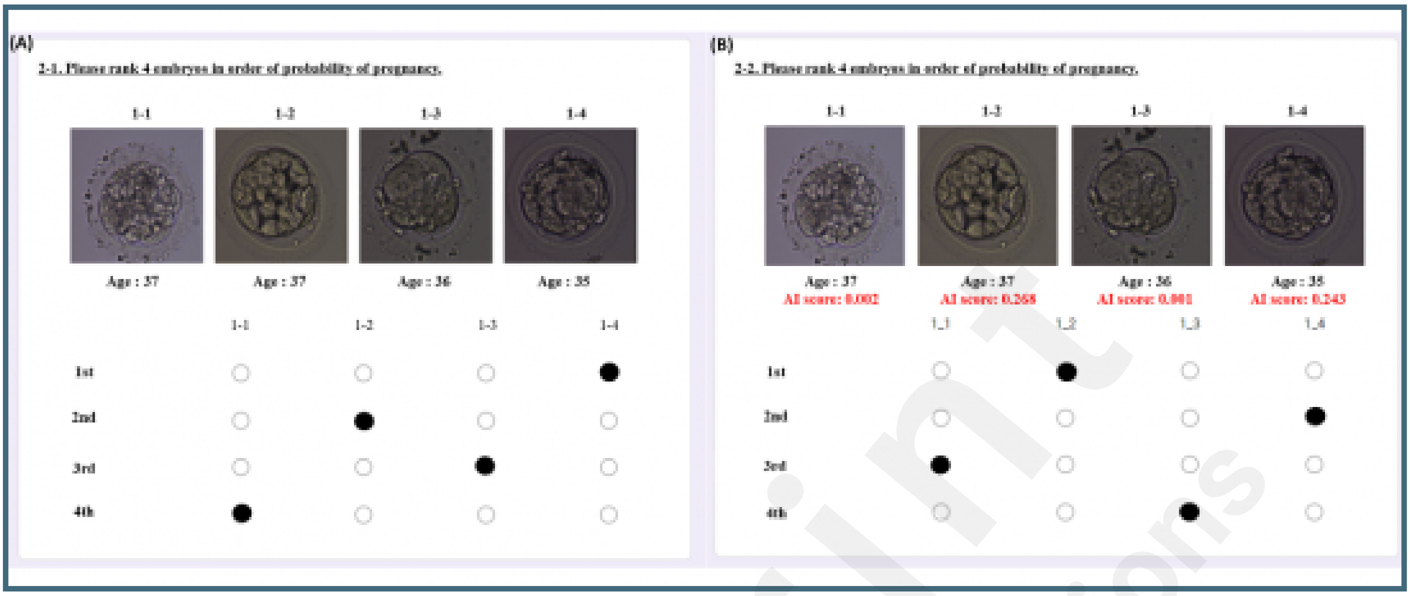
PMID:843571

25. Loewke K, Cho JH, Brumar CD, Maeder-York P, Barash O, Malmsten JE, Zaninovic N, Sakkas D, Miller KA, Levy M, VerMilyea MD. Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos. Fertil Steril The Authors; 2022;117(3):528–535. PMID:34998577
26. Zaninovic N, Irani M, Meseguer M. Assessment of embryo morphology and developmental dynamics by time-lapse microscopy: is there a relation to implantation and ploidy? Fertil Steril Elsevier Inc.; 2017;108(5):722–729. PMID:29101997
27. Athayde Wirka K, Chen AA, Conaghan J, Ivani K, Gvakharia M, Behr B, Suraj V, Tan L, Shen S. Atypical embryo phenotypes identified by time-lapse microscopy: High prevalence and association with embryo development. Fertil Steril Elsevier Inc.; 2014;101(6):1637-1648.e5. PMID:24726214
28. Kirkegaard K, Agerholm IE, Ingerslev HJ. Time-lapse monitoring as a tool for clinical embryo assessment. Hum Reprod Oxford University Press; 2012;27(5):1277–1285. PMID:22419744
29. Aparicio-Ruiz B, Romany L, Meseguer M. Selection of preimplantation embryos using time-lapse microscopy in in vitro fertilization: State of the technology and future directions. Birth Defects Res John Wiley & Sons, Ltd; 2018 May 1;110(8):648–653. PMID:29714056
30. Khan A, Gould S, Salzmann M. Segmentation of developing human embryo in time-lapse microscopy. Proc - Int Symp Biomed Imaging IEEE Computer Society; 2016 Jun 15;2016-June:930–934. doi: 10.1109/ISBI.2016.7493417
31. Chen M, Wei S, Hu J, Yuan J, Liu F. Does time-lapse imaging have favorable results for embryo incubation and selection compared with conventional methods in clinical in vitro fertilization? A meta-analysis and systematic review of randomized controlled trials. PLoS One 2017;12(6). PMID:28570713

Supplementary Files

Figures

Questions on embryo selection from the web survey. (A) Phase 2 of the survey: question on embryo selection without AI scores. (B) Phase 3 of the survey: question on embryo selection with AI scores.



(A)

2-1. Please rank 4 embryos in order of probability of pregnancy.

1-1	1-2	1-3	1-4
Age : 37	Age : 37	Age : 36	Age : 35

	1-1	1-2	1-3	1-4
1st	○	○	○	●
2nd	○	●	○	○
3rd	○	○	●	○
4th	●	○	○	○

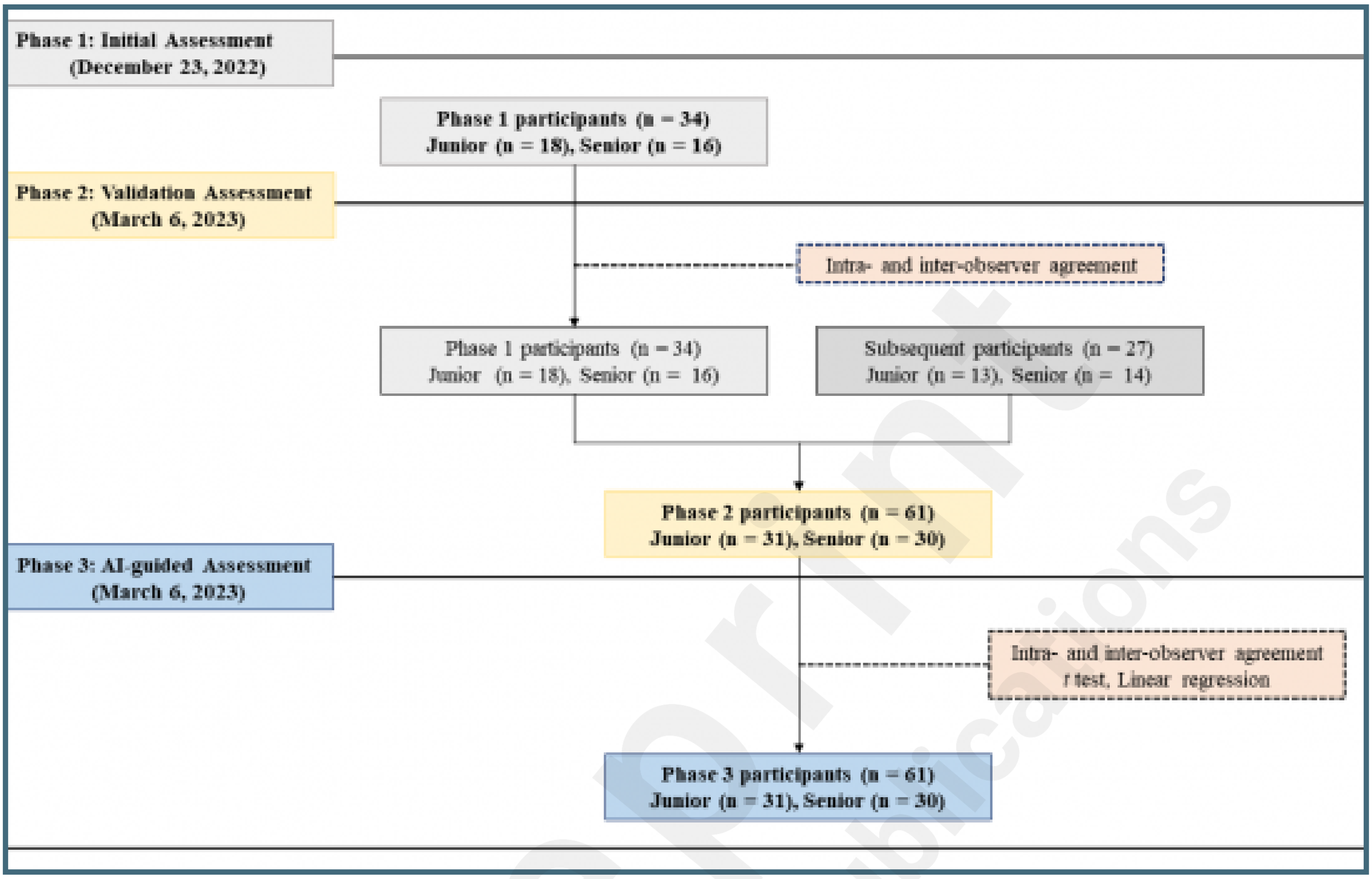
(B)

2-2. Please rank 4 embryos in order of probability of pregnancy.

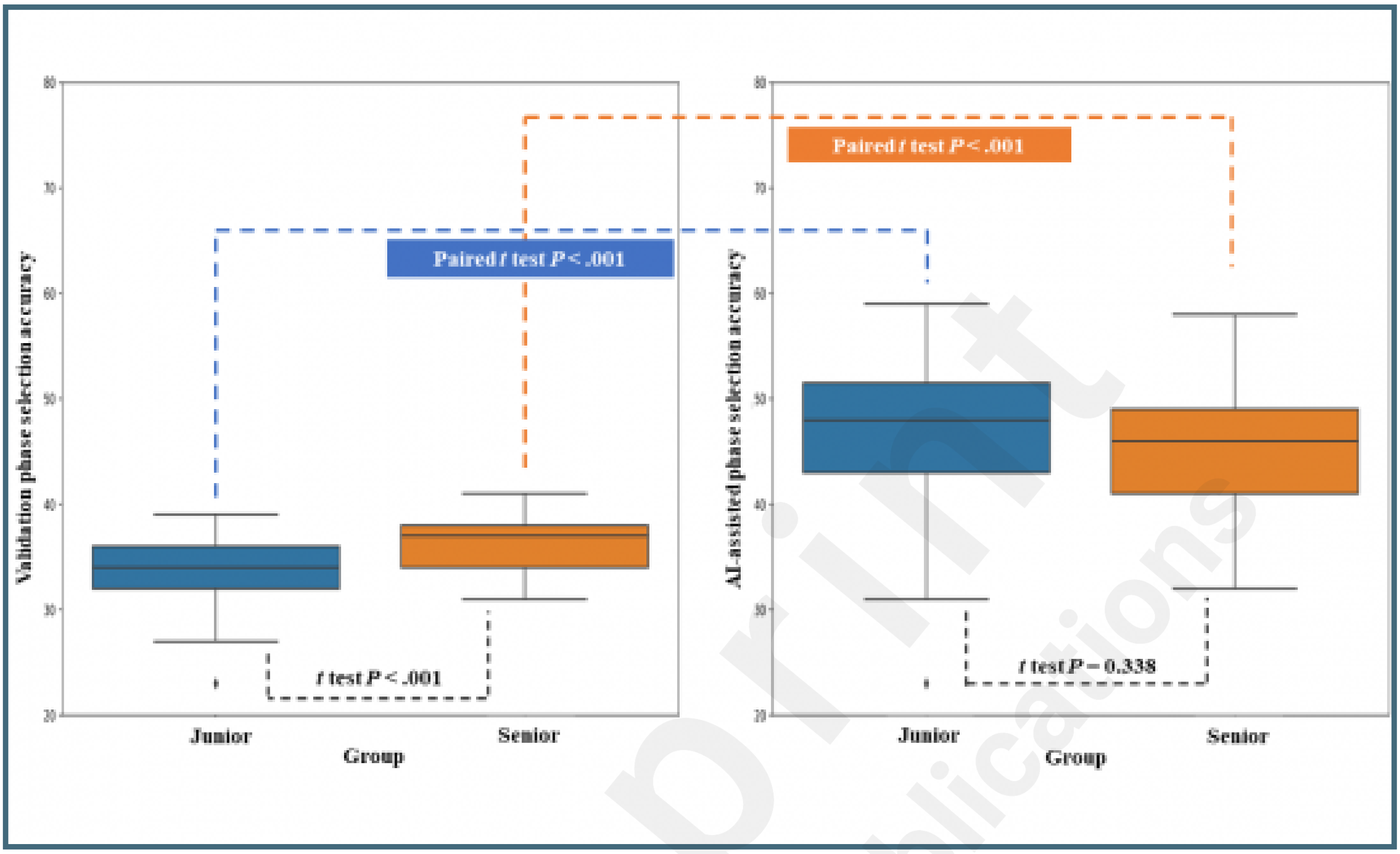
1-1	1-2	1-3	1-4
Age : 37 AI score: 0.002	Age : 37 AI score: 0.268	Age : 36 AI score: 0.001	Age : 35 AI score: 0.243

	1_1	1_2	1_3	1_4
1st	○	●	○	○
2nd	○	○	○	●
3rd	●	○	○	○
4th	○	○	●	○

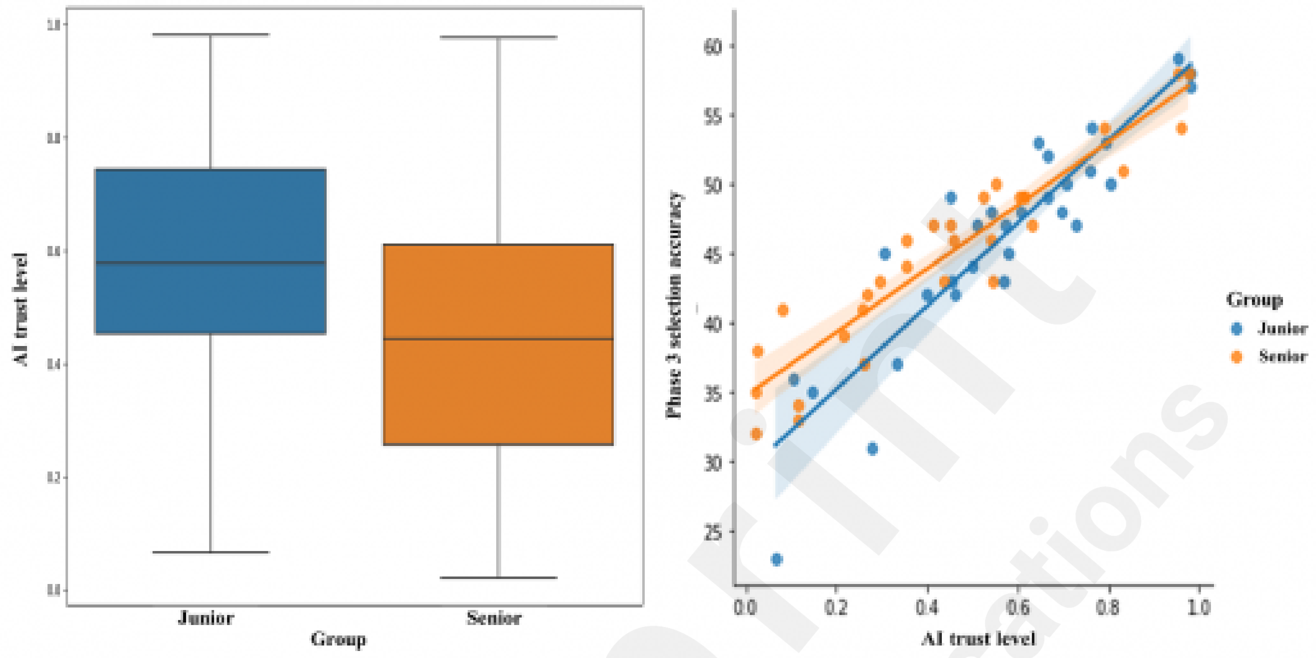
Study design and participant flow diagram according to the survey phase.


Phase 1: Initial Assessment (December 23, 2022)
Phase 1 participants (n = 34) Junior (n = 18), Senior (n = 16)
Phase 2: Validation Assessment (March 6, 2023)
Intra- and inter-observer agreement
Phase 1 participants (n = 34) Junior (n = 18), Senior (n = 16)
Subsequent participants (n = 27) Junior (n = 13), Senior (n = 14)
Phase 2 participants (n = 61) Junior (n = 31), Senior (n = 30)
Phase 3: AI-guided Assessment (March 6, 2023)
Intra- and inter-observer agreement
t test, Linear regression
Phase 3 participants (n = 61) Junior (n = 31), Senior (n = 30)

Within- and between-group t test results.



Results of the groupwise analyses of AI trust level through t test and regression analysis.



Multimedia Appendixes

Distribution of questions in which top-ranked embryos differ between embryologists and AI by group.
URL: http://asset.jmir.pub/assets/431057958d6fccfb798b0a72c5d7e2a9.docx

Correlation between AI scores and traditional manual grading.
URL: http://asset.jmir.pub/assets/b481890966e65087255c34e40b9be043.docx