RBMO

ELSEVIER

ARTICLE

An artificial intelligence model correlated with morphological and genetic features of blastocyst quality improves ranking of viable embryos

BIOGRAPHY

Dr Michelle Perugini is a health technology entrepreneur with extensive experience in healthcare and advanced artificial intelligence technologies. She is an avid supporter of artificial intelligence and women's health sectors globally. Dr Perugini has a PhD in medicine and was a post-doctoral research fellow in oncology for a decade.

Sonya M. Diakiw[1,*], Jonathan M.M. Hall[1,2,3], Matthew VerMilyea[4,5], Adelle Y.X. Lim[6], Wiwat Quangkananurug[7], Sujin Chanchamroen[7], Brandon Bankowski[8], Rebecca Stones[8], Ashleigh Storr[9], Andrew Miller[4], Glen Adaniya[10], RaeAnne van Tol[10], Roberta Hanson[11], Jon Aizpurua[1], Lydia Giardini[12], Adrian Johnston[1], Tuc Van Nguyen[1], Milad A. Dakka[1], Don Perugini[1], Michelle Perugini[1,13]

KEY MESSAGE

Improved methods for evaluating artificial intelligence in the field of IVF are described. The importance of correlating intelligence scores with known parameters of embryo quality for artificial intelligence characterization is highlighted. The findings support artificial intelligence testing methods and use in clinical practice.

ABSTRACT

Research question: Can better methods be developed to evaluate the performance and characteristics of an artificial intelligence model for evaluating the likelihood of clinical pregnancy based on analysis of day-5 blastocyst-stage embryos, such that performance evaluation more closely reflects clinical use in IVF procedures, and correlations with known features of embryo quality are identified?

Design: De-identified images were provided retrospectively or collected prospectively by IVF clinics using the artificial intelligence model in clinical practice. A total of 9359 images were provided by 18 IVF clinics across six countries, from 4709 women who underwent IVF between 2011 and 2021. Main outcome measures included clinical pregnancy

[1] Life Whisperer Diagnostics (a subsidiary of Presagen), San Francisco CA 94404, USA, and Adelaide SA 5000, Australia
[2] Australian Research Council Centre of Excellence for Nanoscale BioPhotonics, Adelaide SA 5000, Australia
[3] School of Physical Sciences, The University of Adelaide, Adelaide SA 5000, Australia
[4] Ovation Fertility, Austin TX 78731, USA
[5] Texas Fertility Center, Austin TX 78731, USA
[6] Alpha IVF and Women's Specialists, Petaling Jaya Selangor 47810, Malaysia
[7] Safe Fertility, Lumpini Bangkok 10330, Thailand
[8] ORM Fertility, Portland OR 97205, USA
[9] Flinders Fertility, Glenelg SA 5045, Australia
[10] Midwest Fertility Specialists, Carmel IN 46032, USA
[11] Ovation Fertility, San Antonio TX 78258, USA
[12] IVF-Spain, Alicante 03540, Spain
[13] Adelaide Medical School, The University of Adelaide, Adelaide SA 5000, Australia

*Corresponding author. E-mail address: sonya@lifewhisperer.com (S.M. Diakiw). https://doi.org/10.1016/j.rbmo.2022.07.018
1472-6483/
Declaration: Life Whisperer Diagnostics, Inc., and Life Whisperer Diagnostics, Pty. Ltd. are wholly owned subsidiaries of the parent company, Presagen Pty. Ltd. Funding for the study was provided by Presagen with grant funding received from the South Australian Government: Research, Commercialisation, and Startup Fund (RCSF).

KEYWORDS

Artificial intelligence
Embryo selection
Gardner score
IVF
PGT-A

outcome (fetal heartbeat at first ultrasound scan), embryo morphology score, and/or pre-implantation genetic testing for aneuploidy (PGT-A) results.

Results: A positive linear correlation of artificial intelligence scores with pregnancy outcomes was found, and up to a 12.2% reduction in time to pregnancy (TTP) was observed when comparing the artificial intelligence model with standard morphological grading methods using a novel simulated cohort ranking method. Artificial intelligence scores were significantly correlated with known morphological features of embryo quality based on the Gardner score, and with previously unknown morphological features associated with embryo ploidy status, including chromosomal abnormalities indicative of severity when considering embryos for transfer during IVF.

Conclusion: Improved methods for evaluating artificial intelligence for embryo selection were developed, and advantages of the artificial intelligence model over current grading approaches were highlighted, strongly supporting the use of the artificial intelligence model in a clinical setting.

INTRODUCTION

Selecting the best quality embryo for transfer is essential to ensure the shortest time to pregnancy (TTP) for patients undergoing IVF procedures. Despite efforts to continually improve outcomes, IVF is by no means a perfected technology, with live birth rates averaging around 30% per embryo transfer. Improved embryo selection methods are a critical driver for increasing pregnancy success rates during IVF.

In recent years, interest in the use of artificial intelligence to support embryo quality assessment has grown, with numerous algorithms being developed for analysis of static images or time-lapse videos of embryos to aid in the selection of embryos for transfer (*Khosravi et al., 2019*; *Tran et al., 2019*; *Chavez-Badiola et al., 2020*; *Silver et al., 2020*; *VerMilyea et al., 2020*; *Berntsen et al., 2022*; *Erlich et al., 2022*; *Loewke et al., 2022*). Because of the infancy of artificial intelligence in the embryology field, however, few of these studies have evaluated real-world clinical use. Furthermore, although artificial intelligence is often compared with existing methods of embryo quality evaluation, such as morphological grading or pre-implantation genetic testing (PGT-A), characterisation of the extent to which artificial intelligence can identify known features of embryo quality is relatively limited.

The evaluation of artificial intelligence for embryo assessment has so far focused on using performance metrics generally suited to binary classification problems. Binary performance metrics do not accurately reflect the clinical application and intended use of artificial intelligence for embryo quality assessment, which is to select an embryo for transfer from a patient cohort containing multiple embryos in any given IVF cycle. These binary performance metrics are, therefore, likely underestimating the power of the artificial intelligence model in this setting.

The Life Whisperer viability artificial intelligence model analyses static images of day-5 blastocyst-stage embryos during IVF procedures to provide information on the likelihood of clinical pregnancy (embryo viability) (*VerMilyea et al., 2020*). In this study, the viability artificial intelligence model was characterized using data collected prospectively through the course of real-world clinical use. This included development of a novel method for performance evaluation that assesses the ability of the artificial intelligence model to rank embryos within a patient cohort during a single IVF cycle to select the best embryo for transfer.

Results showed a strong correlation of artificial intelligence scores with known embryo quality measures, including those measured by Gardner-based morphological grading and PGT-A, and demonstrated superior embryo ranking performance for the viability artificial intelligence model. This study provides insight into the mechanisms of artificial intelligence-based evaluation of embryos and substantiates the use of artificial intelligence for embryo viability assessment in clinical use for the selection of embryos during IVF procedures.

MATERIALS AND METHODS

Artificial intelligence model development

The development of the viability artificial intelligence model was described by *VerMilyea et al. (2020)*. In brief, the artificial intelligence model consists of a series of image pre-processing steps, following by object detection, optional segmentation of the zona pellucida or intra-zonal cavity region, and a selection of image-based deep learning classification models, which are combined together in an ensemble model that provides a final artificial intelligence score for each image. The pre-processing steps include the following: alpha-channel stripping, colour normalization, square-padding and cropping. The deep learning classification models comprise ResNet and DenseNet convolutional neural network architectures, which were trained by minimizing the log loss using Stochastic Gradient Descent. Images used for training comprised a wide range of specifications, e.g. resolutions, brightness, contrast, colour balance, and were obtained from multiple different imaging systems to ensure the resulting model was robust to input variation. In addition, augmentation of images was used during training to anticipate changes to lighting conditions, rotation of the embryo and focal length, e.g. rotations, reflections, Gaussian blur, contrast variation and random compression. No additional patient meta-data were included in developing the artificial intelligence model.

Experimental design

For the present study, de-identified data were provided for a total of 4709 women aged 18 years or over who underwent IVF procedures at 18 different IVF clinics between 2011 and 2021. Data included 9359 two-dimensional embryo images with associated clinical outcomes, including clinical pregnancy outcome, preimplantation genetic testing with aneuploidy (PGT-A) outcome and Gardner score. A description of the test datasets is provided in Supplementary Tables 1–3. The full dataset is provided in Supplementary Table 4.

For inclusion in the study, images of embryos were required to be taken on day 5 of culture post-fertilization using either standard or time-lapse-based optical light microscopy systems. Images were excluded if taken after biopsy for PGT-A or cryopreservation. All images were required to have a minimum resolution of 480 × 480 pixels with the complete embryo in the field of view, and the focal plane centred on the inner cell mass.

For evaluating performance in predicting clinical pregnancy outcomes, data were limited to patients who received a single embryo transfer with a day 5 blastocyst-stage embryo. For evaluating the correlation of artificial intelligence scores with the Gardner system, data were limited to embryo images with an associated embryologist's grade supplied. For evaluating the correlation of artificial intelligence scores with embryo ploidy status, data were limited to embryo images with associated results obtained via conventional PGT-A methods.

Collection of retrospective data for this study was exempted from ethical review and approval, and from the requirement for informed consent, because of the retrospective nature of the analyses, and de-identification of data. Exemption was confirmed by Sterling IRB ID number 6467 (5 September 2018) and number 7751 (21 January 2020), for protocol identifiers LW-C-001A and LW-C-004A, respectively. Collection of prospective data for this study was carried out in accordance with the Life Whisperer patient privacy policy, constituting informed consent for research purposes. This study was conducted according to the guidelines of the Declaration of Helsinki of 1975, as amended.

Embryo scoring methods

The artificial intelligence score is presented on a scale of 0.0 to 10.0, with 10.0 representing the highest likelihood of clinical pregnancy (highest confidence that the embryo is viable) and 0.0 representing the least likelihood of clinical pregnancy (highest confidence that the embryo is not viable). A viable embryo was defined as an embryo for which at least one fetal heartbeat was detected on first ultrasound scan at around 6 weeks after transfer, and a non-viable embryo was defined as one for which no fetal heartbeat was detected. For live birth analyses, a viable embryo was defined as one for which a successful live birth occurred, and a non-viable embryo was defined as one for which no live birth occurred.

For correlation of artificial intelligence score with known morphological features of embryo quality, embryo images were graded according to their developmental stage as first described by *Gardner and Schoolcraft (1999)* (the Gardner scoring system). Embryo grades were provided by the treating embryologist at the time the image was taken.

Preimplantation genetic testing for aneuploidy

The present study used retrospective PGT-A results carried out by Ovation Fertility Genetics Laboratories. Testing was carried out on biopsies from day 5 blastocyst-stage embryos using a next-generation sequencing-based assay (VeriSeq platform) (Illumina, San Diego, CA). Testing followed standard protocols and manufacturer recommendations.

Resulting chromosome analyses were provided using standard cytogenetic nomenclature to record abnormalities. Embryos with 30% abnormal cells or less were classified as euploid, and those with over 70% cells were classified as aneuploid. Embryos were defined as mosaic if they contained between 31% and 70% abnormal cells.

Statistical analyses

This study involved standard statistical methods used in performance evaluation of machine learning classifiers, including accuracy and area under the curve (AUC) value for receiver operating characteristic curves (ROC) (*Florkowski, 2008*). Average artificial intelligence scores for multiple groups were compared using ordinary one-way analysis of variance with Tukey's multiple comparisons post-test. Trends in average artificial intelligence scores were evaluated using ordinary one-way analysis of variance with test for trend between column mean and left-to-right column order, unless otherwise indicated. Trends in the proportion of successful pregnancies or the proportion of euploid embryos were evaluated using a chi-squared test for trend. Error bars indicate standard error of the mean where presented, and $P < 0.05$ was considered significant.

A five-point moving average method was used to evaluate the correlation between every point on the artificial intelligence score scale, e.g. 0.0, 0.1, 0.2 through to 10.0 out of 10.0, and the proportion of successful clinical pregnancies. This involved taking the proportion of viable embryos at the point of interest plus two points above and two points below (five-point total). GraphPad Prism version 9.0.0 was used for statistical analyses.

Simulated cohort ranking analyses

For simulated cohort ranking analyses, images of transferred embryos were randomized into tens of thousands of simulated embryo 'cohorts' each consisting of embryos with known pregnancy outcomes from different patients. The number of distinct cohorts generated is related to dataset size and cohort size as follows: *n* cohorts ≈ *n* images in dataset / (average) cohort size * 1000. To construct the simulation, embryo images within a dataset were compiled into a single list and shuffled randomly. Then, starting from the top of the list, the first *n* images were taken to be a simulated cohort (depending on the cohort size allocated for that experiment). These embryo images were then removed from the list, and the process repeated until the value of *n* was larger than the remaining embryos in the list (remaining embryos were discarded). Cohorts containing all viable or all non-viable embryos were then removed from the analysis. This procedure constituted a single randomization of the dataset, which was carried out for a pre-set random seed to ensure repeatability. To obtain additional randomizations, the random seed was incremented, and the process repeated. For each experiment, a total of 1000 randomizations were generated, and the results of all randomizations analysed together. Therefore, each embryo in a dataset was used in up to 1000 discrete cohorts for each experiment, but in each case the embryo was most likely allocated to cohorts of different compositions (including being grouped with different neighbouring embryos and, where relevant, in cohorts of different sizes). After the randomization of embryos to simulated cohorts was completed, each cohort was sorted in a descending order rank score. Rank scores used in these experiments were either artificial intelligence model scores (on a scale of 0.0 to 10.0), the randomly generated scores used for cohort randomization (as a proxy for embryologists' ranking based on pregnancy success rates of about 50%), or a Gardner-based rank

score. Three independent Gardner-based ranking methods were investigated in the present study. The first used a single binary threshold of 3BB as a commonly reported example of embryo stratification (*Kemper et al., 2021*). The second used a four-group ranking system as follows: rank 1 = 3-6AA; rank 2 = 3-6AB/BA, 2AA; rank 3 = 3-6BB/AC/CA, 2AB/BA; and rank 4 = 2-6BC/CB/CC, 2BB/AC/CA, 1XX. This four-group ranking system has been used in several studies to define embryo groups as very good, good, fair and poor quality (*Capalbo et al., 2014*; *Irani et al., 2017*; *Zhao et al., 2018*). The final method was a novel seven-group ranking system defined in the present study using published research, using the following general rules: expansion grades 3-6 > grade 2 > grade 1 and combined inner cell mass and trophectoderm grades AA > AB/BA > BB > AC/CA > BC/CB > CC. The rank groups for the seven-group system were as follows: rank 1 = 3-6AA; rank 2 = 3-6AB/BA; rank 3 = 3-6BB; rank 4 = 3-6AC/CA, 2AA; rank 5 = 3-6BC/CB; 2AB/BA; rank 6 = 3-6CC, 2BB; and rank 7 = 2AC/CA/BC/CB/CC, 1XX.

For a single simulated cohort, theoretical TTP or time to live birth was calculated as the position of the top viable embryo in the ranked cohort. Mean difference in TTP or time to live birth was compared for all cohorts between ranking methods using the asymmetric Laplace distribution (ALD) probability density function (*Kozubowski and Podgorski, 2000*). Significance was determined using the Mann–Whitney Wilcoxon U test. For estimation of theoretical pregnancy rates or live birth rates, the percentage of simulated cohorts where the top-ranked embryo was viable was calculated for all ranking methods. This value provides a representation of pregnancy rate or live birth rate for the first cycle of IVF. Significant differences were determined using a chi-squared test.

RESULTS

Methods evaluating the correlation of artificial intelligence score with clinical pregnancy rate to assess relative likelihood of embryo viability

The viability artificial intelligence model was previously evaluated for its ability to predict clinical pregnancy using a binary prediction threshold of 5.0/10.0. In the study by *VerMilyea et al. (2020)*, the accuracy of the artificial intelligence model was 64.3%, and the ROC-AUC was 0.68. These values, while allowing a general comparison to be made with similar reported artificial intelligence models, e.g. Alife Health artificial intelligence model ROC-AUC 0.62–0.64, Fairtility artificial intelligence model ROC-AUC 0.68–0.70, Vitrolife artificial intelligence model ROC-AUC 0.67 (*Berntsen et al., 2022*; *Erlich et al., 2022*; *Loewke et al., 2022*), do not give an indication of the ability of the artificial intelligence model to effectively rank embryos for selection within a patient cohort during a single IVF cycle.

To evaluate the performance of the viability artificial intelligence model in a way that is similar to its intended clinical use, a number of additional methods were developed to evaluate its ability to rank embryos according to their likelihood of clinical pregnancy.

For development of these ranking methods, a retrospective blind test dataset consisting of 1161 day 5 embryo images with corresponding clinical pregnancy outcomes was used (Supplementary Table 1). Images were provided by 1158 women who underwent IVF between 2011 and 2018 in one of 14 IVF clinics across the USA, Australia, New Zealand, Malaysia and Thailand. Results of conventional performance evaluation on this dataset using a binary prediction threshold of 5.0/10.0 were similar to the findings described by *VerMilyea et al. (2020)* (accuracy of 61.8% and ROC-AUC of 0.61 in the present study).

The first two methods that were developed to assess ranking ability used correlative analyses of artificial intelligence score and pregnancy rate to reflect the relative likelihood of embryo viability on transfer. The simplest of these methods involved defining four 'likelihood categories' (bins) corresponding to different artificial intelligence score thresholds, then evaluating the proportion of embryos resulting in successful clinical pregnancies for each category. Results showed a significant positive correlation between categories of increasing artificial intelligence score and the proportion of successful pregnancies using a chi-squared test for trend (FIGURE 1A), with the percentage of pregnancies more than doubling from 29.3% in the low likelihood category to 64.6% in the very high likelihood category ($P < 0.0001$). Therefore, it can be inferred that embryos in the highest score category are twice as likely to result in pregnancy as those in the lowest score category. This basic type

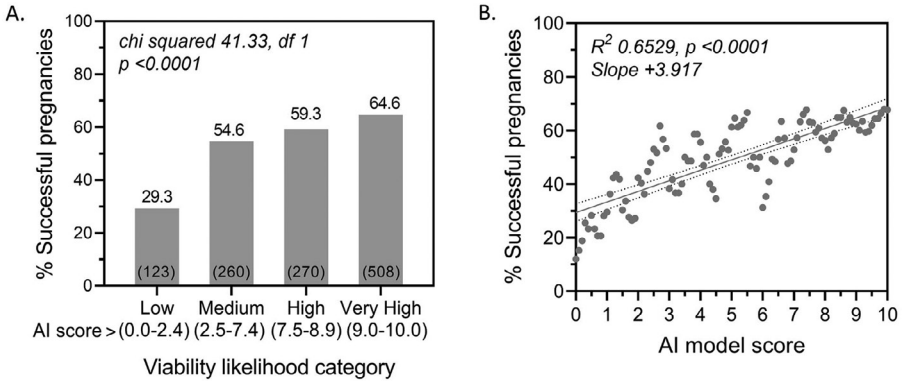

FIGURE 1 Methods evaluating the correlation of artificial intelligence score with clinical pregnancy rate. (A) The correlation between artificial intelligence score and the proportion of successful pregnancies was evaluated using four defined viability likelihood categories based on artificial intelligence score brackets as indicated. The pregnancy rate is depicted for each group, and the number of embryos in each group indicated in parentheses; (B) the linear correlation between each point of the artificial intelligence score and the proportion of successful pregnancies was evaluated using a five-point moving average method of artificial intelligence. AI, artificial intelligence; df, degrees of freedom.

of analysis can be easily conducted within individual IVF laboratories to validate artificial intelligence model performance on relevant internal datasets.

To evaluate the correlation between artificial intelligence score and the proportion of successful clinical pregnancies at a higher level of granularity, the linearity of the association was assessed using a five-point moving average method across every point of the artificial intelligence scale (from 0.0 to 10.0). Linear regression analysis (FIGURE 1B) demonstrated a significant linear correlation between the artificial intelligence score and the proportion of successful pregnancies ($P <$ 0.0001), with a slope of approximately +4, and values ranging from 12.0% viable embryos at the lowest point (0.0) to 67.7% viable embryos at the highest point (10.0) of the artificial intelligence scale. The observed positive correlation reflects the increased likelihood that embryos of higher scores will lead to pregnancy over embryos of lower scores along the artificial intelligence score scale. The apparent oscillatory behaviour of the five-point evaluation is an inherent characteristic of the moving average method in general, which introduces an artificial wave effect. This effect is expected to be reduced as dataset size increases.

Development of a simulated cohort ranking analysis method to evaluate relative ranking ability using theoretical time to pregnancy and first-cycle pregnancy rate

The final method developed to evaluate artificial intelligence ranking ability was a sophisticated ranking analysis using simulated patient 'cohorts' to estimate the number of cycles that would be needed to select a viable embryo leading to pregnancy, with results presented as theoretical TTP, and the proportion of cohorts with a top-ranked viable embryo, with results presented as theoretical pregnancy rate for one cycle of IVF. For this analysis, the 1161 images in the viability test dataset were randomized multiple times to give a total of approximately 110,000 distinct simulated patient embryo cohorts of an average of approximately 10 embryos each. Each cohort was intended to represent a single patient cohort during IVF, but where pregnancy outcomes were known for every embryo. This method was developed to address one of the greatest obstacles facing the evaluation of artificial intelligence for embryo selection, as the pregnancy outcome of every embryo in a real-world patient cohort would not generally be known, limiting the ability to conduct this type of analysis. Using this approach, the embryos in each simulated cohort were ranked according to the likelihood of clinical pregnancy based on artificial intelligence score. Time to pregnancy was calculated as the position of the first embryo in the ranked cohort to result in clinical pregnancy, and first-cycle pregnancy rate was calculated as the percentage of cohorts with a viable embryo in the top-ranked position. The method is presented in FIGURE 2A.

For this initial analysis, the distribution of cohort sizes generated was based on a real-world clinical dataset of 487 patient IVF cycles. The average size of all cohorts in this dataset was 9.3 embryos per cohort; however, given the simulated cohort ranking analyses of necessity excludes cohorts with no embryos (failed IVF cycles), the average size excluding these failed IVF cycles was 9.7 embryos. The distribution of cohort sizes in this real-world dataset compared with the final distribution of cohort sizes generated in the simulated cohort ranking experiment is presented in FIGURE 2B. The distribution of cohort sizes according to the percentage of viable embryos (successful pregnancies) is also presented. Although a large proportion of cohorts demonstrated a 1:1 ratio of viable to non-viable embryos, the distribution itself demonstrated a general shift to a slightly higher percentage of viable embryos (average 52.3%), owing to the marginal bias towards viable embryos in the original dataset (57.4% viable).

The distributions of TTP values for artificial intelligence ranking, random ranking and the differences between them are presented in FIGURE 2C. The TTP distributions showed exponential decay behaviour consistent with the expected probabilities for correctly identifying a viable embryo. The TTP pregnancy values for the artificial intelligence model were more weighted towards achieving the best possible TTP value of 1 compared with random ranking, whereas random ranking gave rise to more TTPs of >1. This indicated that the model systematically provided viable embryos with higher scores than those achieved by random ranking. The probability distribution of the difference of these two exponential decay curves (with different exponents) was best modelled using an ALD (FIGURE 2C). The shape of the distribution was asymmetrically biased towards negative values, showing that the artificial intelligence model was generally reducing TTP values relative to random ranking.

The mean TTP using artificial intelligence ranking for all of the approximately 110,000 simulated cohorts was 1.504 cycles, representing a 12.6% ± 0.3% reduction in the number of cycles needed to select a viable embryo compared with random ranking (mean TTP of 1.720 cycles). The statistical uncertainty associated with the mean TTP was calculated as 0.003 cycles for both the artificial intelligence model and random ranking TTP values using the ALD probability function ($P <$ 0.0001). In addition, pregnancy rates of 66.0% and 56.5% were observed for artificial intelligence ranking and random ranking, respectively, representing an improvement of 9.5% in theoretical pregnancy rate for one cycle of IVF ($P < 0.0001$). These improvements in performance are significant, given that all the embryos in the ranking study were selected for transfer and, therefore, represent a bias towards a higher quality of embryo.

Although the distribution of embryo cohort sizes used in this analysis was derived from a real-world clinical dataset, average cohort sizes can vary significantly between clinics based on a variety of factors, such as patient age and medical history. An investigation of the contribution of cohort size to the results obtained using simulated cohort ranking analyses was carried out using fixed cohort sizes ranging from 3 to 12 embryos. Results showed that TTP and pregnancy rate generally improved as cohort sizes increased, for both the artificial intelligence model and random ranking. The observed improvement, however, was greater for the artificial intelligence model than random ranking, resulting in a net increase in improvement with increasing cohort size. The artificial intelligence model demonstrated improvements over random ranking, ranging from an 8.6% to a 14.2% reduction in TTP, and a 9.2% to a 10.4% improvement in first-cycle pregnancy rate, for cohort sizes of 3 to 12, respectively. Performance values for all cohort sizes are presented in TABLE 1.

Although clinical pregnancy is one measure of IVF success, the goal of the

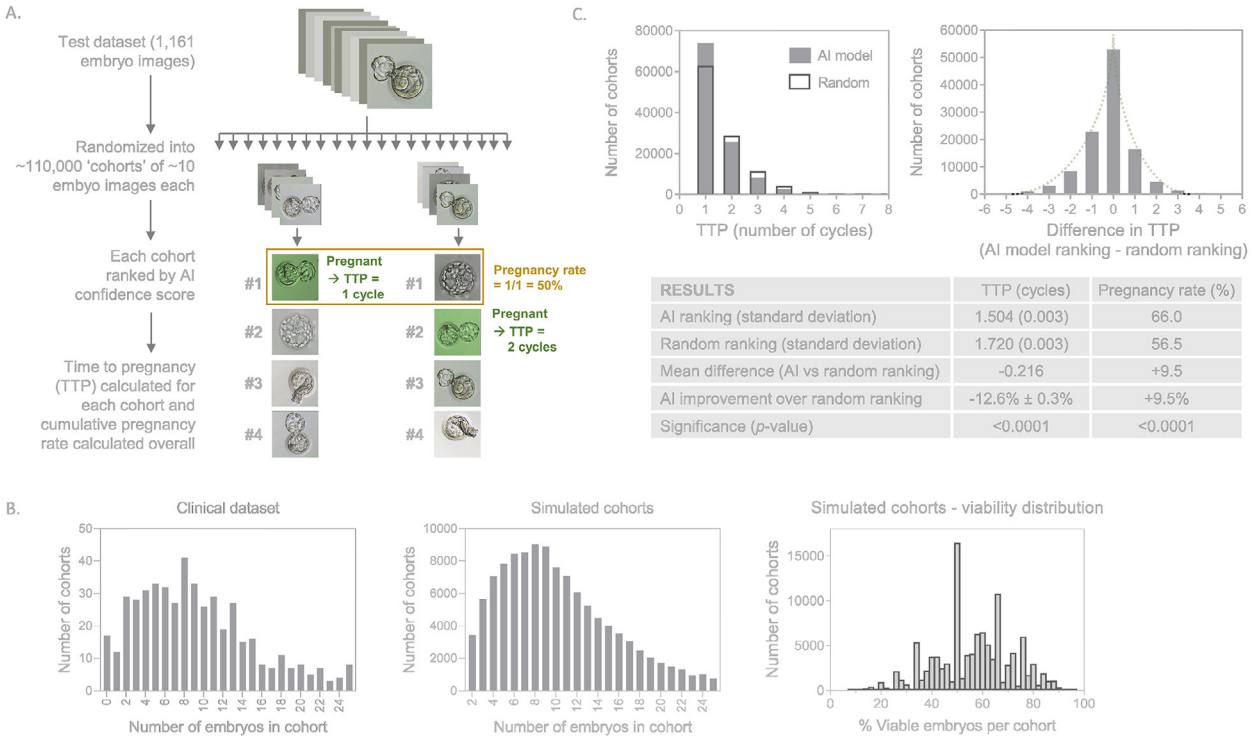

FIGURE 2 Simulated cohort analyses evaluating artificial intelligence ranking ability based on the likelihood of clinical pregnancy. (A) The ability of the artificial intelligence model to rank embryos within a patient cohort and improve theoretical time to pregnancy (TTP) (in number of cycles) and theoretical first-cycle pregnancy rate was evaluated using a novel method using multiple randomized simulated patient cohorts. Embryo images from the viability test dataset were randomized to approximately 110,000 distinct patient 'cohorts' consisting of embryos from different patients, where the clinical pregnancy outcome was known for each embryo. Ranking ability was evaluated by estimating TTP based on the number of cycles needed to select a viable embryo leading to clinical pregnancy, and by estimating pregnancy rate for one cycle of IVF based on the proportion of cohorts with a viable embryo in the top-ranked position; (B) the distributions of cohort sizes in the original clinical dataset used to determine embryo cohort size range, and in the simulated cohort analysis. The distribution of cohorts according to the proportion of viable embryos in each is also depicted for the simulated analysis; (C) results of the simulated cohort analysis are presented for the retrospective viability test dataset. The distributions of TTP values for the artificial intelligence model and random ranking are shown for comparison, as is the distribution of the differences in TTP values, which displays the characteristic asymmetric Laplace distribution indicated by a dotted line. Note that x-axes for the TTP figures are truncated for optimal visualization of distributions; this does not affect interpretation of results (approximately 50–100 cohorts from a total of approximately 110,000 out of range).

TABLE 1 PERFORMANCE VALUES GENERATED USING SIMULATED COHORT RANKING ANALYSES

	Time to pregnancy			First-cycle pregnancy rate		
	Artificial intelligence model (cycles)	Random (cycles)	Artificial intelligence improvement over random (%)[a]	Artificial intelligence model (%)	Random (%)	Artificial intelligence improvement over random (%)[b]
Cohort size						
Clinical distribution[c]	1.504	1.720	12.6	66.0	56.5	9.5
Three embryos	1.476	1.616	8.6	61.8	52.6	9.2
Six embryos	1.531	1.733	11.7	65.4	56.1	9.3
10 embryos	1.506	1.743	13.6	67.2	57.1	10.1
12 embryos	1.495	1.743	14.2	67.6	57.2	10.4
Outcome[d]						
Pregnancy	1.416	2.121	33.2	71.7	45.2	26.5
Live birth	1.686	2.252	25.1	57.3	42.3	15.0

[a] Improvement in time to pregnancy is presented as relative % improvement.

[b] Improvement in first-cycle pregnancy rate is presented as absolute % improvement.

[c] The distribution of embryo cohort sizes was based on a real-world clinical dataset (FIGURE 2B).

[d] Differences in outcome measures were evaluated on a subset of 155 embryos from the total retrospective viability dataset (Supplementary Table 1).

[e]For this analysis only performance values represent time to live birth and first-cycle live birth rate.

procedure is to achieve a successful live birth. To investigate the ability of the artificial intelligence model to rank embryos according to the likelihood of live birth, simulated cohort ranking analyses were carried out on a subset of embryos from the retrospective viability dataset for which live birth outcomes were available (n = 155). Results showed that the artificial intelligence model was in fact able to improve time to live birth over random ranking (P < 0.0001), although this improvement was less than the improvement in time to pregnancy (25.1 ± 0.810% and 33.2 ± 0.693% improvement for time to live birth and pregnancy, respectively) (TABLE 1). Similarly, the artificial intelligence model showed improvement in first-cycle live birth rate over random ranking (+15.0%), although this was again lower than the improvement in first-cycle pregnancy rate (+26.5%). Although these results should be interpreted with caution owing to the smaller dataset size for which live birth outcomes were available, they are supportive of the ability of the artificial intelligence model to rank embryos according to the likelihood of live birth outcome.

Artificial intelligence scores correlate with morphological indicators of embryo quality and show improved embryo ranking in a prospectively collected real-world clinical dataset

To evaluate the correlation betwen artificial intelligence scores and known visible features of blastocyst morphology, a dataset of 2729 day 5 embryo images with matched Gardner scores was prospectively collected during real-world clinical use by two IVF clinics (Spain and Australia). Neither clinic had previously provided any data for training or testing the artificial intelligence model, and therefore, this test dataset was completely independent of model development (Supplementary Table 2). All images were taken using time-lapse imaging systems (Vitrolife EmbryoScope or Merck GERI) for 979 women who underwent IVF procedures between 2019 and 2021.

There is some evidence that the Gardner score and its relative morphological components correlate with pregnancy rates and other measures of clinical outcome (*Gardner and Balaban, 2016*). Given that the artificial intelligence model in this study was trained to evaluate the likelihood of clinical pregnancy, it was hypothesized that the artificial intelligent score would also correlate with components of the Gardner score. Using the Gardner test dataset, the average artificial intelligence score was evaluated for the relative grades of each component, including expansion grade, inner cell mass grade and trophectoderm grade (FIGURE 3A). A significant correlation was found between artificial intelligence score and grade in each case, with the average score generally increasing with advancing blastocyst developmental stage and increasing quality. Similar correlations were observed in the proportion of successful pregnancies in each group (for the subset of images with pregnancy outcomes) (FIGURE 3B).

Although the Gardner scoring method has long given embryologists a valuable way to quantify the physical features of embryos associated with embryo development and quality, it is difficult to use for embryo ranking as it is not intended as a continuous linear scale. Embryologists use different thresholds to define a good versus poor-quality embryo, such as 2AA, 3BB or 3AA (*Gardner et al., 2000*; *Hara et al., 2005*; *Gardner et al., 2007*; *Kemper et al., 2021*); however, no consensus has been reached to date on how to use the Gardner method to rank embryos within a cohort. To compare ranking performance, three distinct Gardner ranking methods were used, including a commonly reported binary 3BB threshold, a previously published four-group ranking system and a novel seven-group ranking system based on Gardner scores that was defined using published research. Each of these grouping methods demonstrated a correlation between morphological rank group and the proportion of successful clinical pregnancies (FIGURE 3C).

These ranking methods were used to evaluate TTP and pregnancy rates on the subset of images from the Gardner dataset that had pregnancy outcomes, allowing for a direct comparison of the ranking abilities of the artificial intelligence model with the Gardner scoring systems. The 547 images with pregnancy outcomes were randomized to approximately 50,000 distinct cohorts for evaluation. Distributions of cohort sizes, TTP values and differences in TTP values are presented in Supplementary Figure 1. For this dataset, the artificial intelligence model demonstrated a 17.9% ± 0.5% improvement over random ranking (TABLE 2). All morphological ranking methods also showed a significant (P < 0.0001) reduction in TTP compared with random ranking, although the improvement was less than that of the artificial intelligence model in each case (6.5% to 10.2% improvement, ± 0.5% assuming independent random bootstraps). It was noted that the ranking ability of the Gardner score improved from the single binary 3BB threshold to the four-group ranking system, then only marginally from the four-group system to the seven-group ranking system defined in this study. These TTP results suggest that only minor improvements may be gained by introducing further granularity to Gardner-based ranking methods, although evaluation of larger datasets including more embryos of poorer morphological quality are warranted.

Depending on the morphological ranking method used, the artificial intelligence model demonstrated improvements of between 8.6% and 12.2% ± 0.5% over the Gardner score in reducing TTP (P < 0.0001 for all comparisons). Consistent with a reduced TTP, the artificial intelligence model also demonstrated improvements in first-cycle pregnancy rate of 9.2% to 11.1% over the various morphology-based ranking methods (P < 0.0001). These results collectively suggest that the artificial intelligence model is superior in its ability to select viable embryos over and above the information provided by the Gardner score.

It was of interest to note that performance values for the artificial intelligence model and random ranking varied between the original retrospective viability dataset (Supplementary Table 1) and the Gardner dataset (Supplementary Table 2), with the artificial intelligence model providing an improvement in TTP of 12.6% and 17.9% over random ranking, respectively. While these datasets differ in size and, therefore, in the number of simulated cohorts generated for each analysis (approximately 110,000 and 50,000, respectively), it was hypothesized that the main factor contributing to this variation was the composition of each dataset in terms of the proportion of successful clinical pregnancies. To investigate this further, the retrospective viability dataset and Gardner dataset were both balanced by removing a random selection of viable embryos in the case of the retrospective viability dataset (n = 171), and a random selection of non-viable embryos in the case of the Gardner dataset (n = 27). Simulated cohort ranking analyses

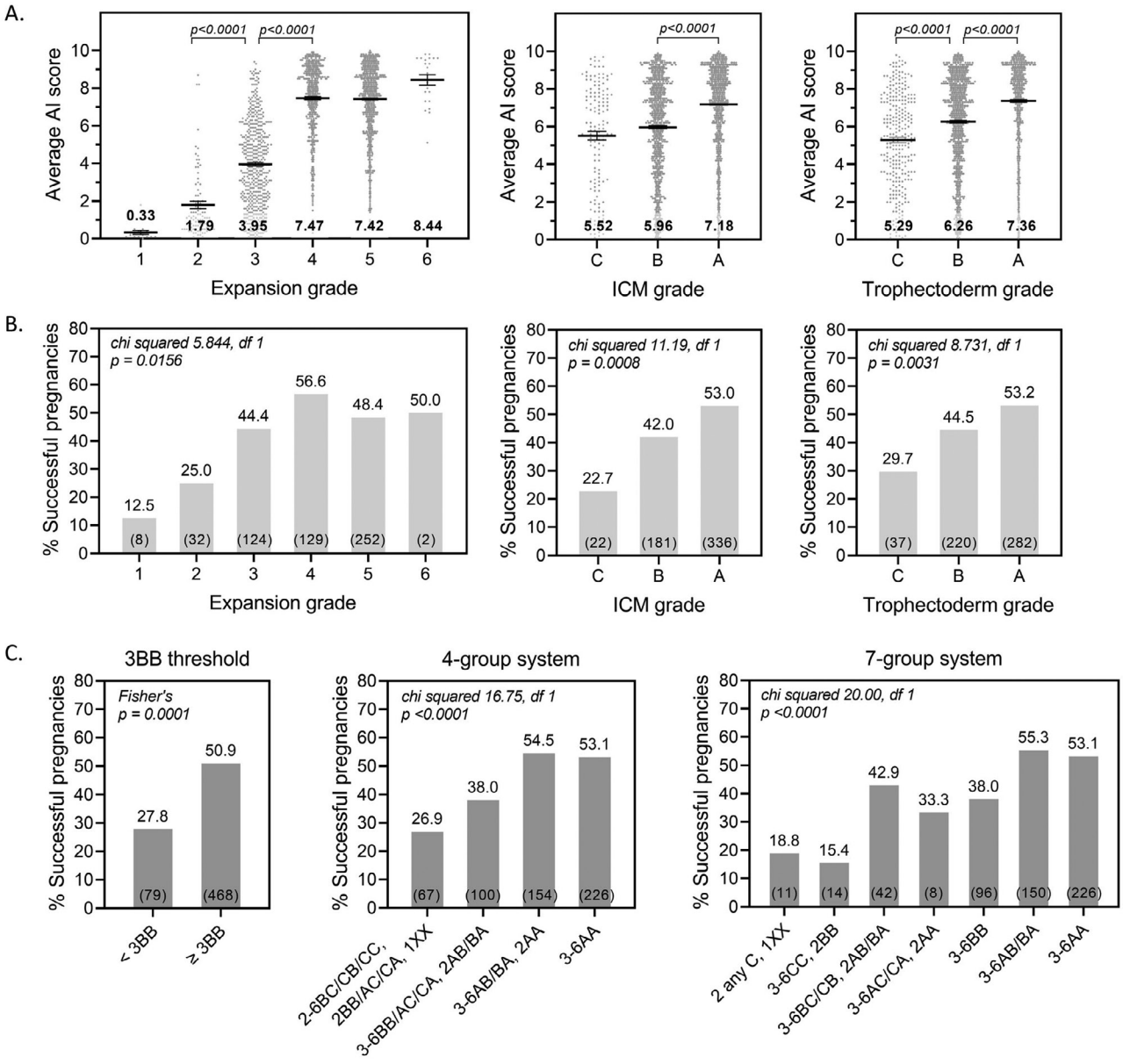

FIGURE 3 Significant correlations between the artificial intelligence score, morphological components of the embryo, and clinical pregnancy outcome. (A) Correlations between average artificial intelligence score and each individual morphological component according to the Gardner score, including expansion grade, inner cell mass (ICM) grade and trophectoderm grade. Average artificial intelligence scores are depicted for each group, and *P*-values shown between adjacent grades only (analysis of variance $P < 0.0001$ for each component); (B) the proportion of successful clinical pregnancies associated with each component of the Gardner score. Pregnancy rate is depicted for each group, and the number of embryos in each group indicated in parentheses; (C) the proportion of successful clinical pregnancies in each rank group according to Gardner score ranking methods. Rank groups were defined using a 3BB threshold, a four-group system as indicated and a seven-group system as indicated. Pregnancy rate is depicted for each group, and the number of embryos in each group indicated in parentheses. AI, artificial intelligence; df, degrees of freedom.

conducted on these balanced datasets demonstrated first-cycle pregnancy rates of 50% each, as would be expected for random selection from a dataset consisting of a 1:1 ratio of viable to non-viable embryos (TABLE 2). The TTPs, while extremely similar in both datasets (1.932 and 1.924), were slightly less than the 2.0 value that would be expected in this situation for random ranking (as in a coin toss experiment). This was the result of introducing experimental conditions, such as excluding cohorts with no viable or non-viable embryos, the proportion of which differ between datasets based on the overall ratio of viable to non-viable embryos. In estimation of TTP, it was necessary to exclude these cohorts, because as well as providing no further constructive information to interpretation of the results (the effect on performance values would be the same for both random ranking and artificial intelligence ranking), it would also be impossible to assign a TTP value to cohorts containing no viable embryos using this approach as that cohort would never lead to pregnancy.

Although it is not possible to evaluate TTP for cohorts with no viable embryos, the ability of the artificial intelligence model to identify whether a cohort contained at least one viable embryo versus no viable embryos was assessed using a modified version of the simulated

TABLE 2 COMPARISON OF THE RANKING ABILITY OF THE ARTIFICIAL INTELLIGENCE MODEL WITH GARDNER-BASED RANKING METHODS

Assessment	Raw values			% Improvement[a]		
	Artificial intelligence model	Random	Gardner	Artificial intelligence versus random	Gardner versus random	Artificial intelligence versus Gardner
Time to pregnancy[b]						
3BB Gardner threshold	1.651	2.012	1.881	17.9	6.5	12.2
Four-group Gardner ranking	1.651	2.012	1.809	17.9	10.1	8.7
Seven-group Gardner ranking	1.651	2.012	1.806	17.9	10.2	8.6
Balanced Gardner dataset	1.618	1.924	ND	15.9	ND	ND
Balanced retrospective viability dataset	1.678	1.932	ND	13.2	ND	ND
First-cycle pregnancy rate						
3BB Gardner threshold	62.6	47.9	51.5	14.7	3.6	11.1
Four-group Gardner ranking	62.6	47.9	53.4	14.7	5.5	9.2
Seven-group Gardner ranking	62.6	47.9	53.4	14.7	5.5	9.2
Balanced Gardner dataset	63.3	50.1	ND	13.2	ND	ND
Balanced retrospective viability dataset	58.8	50.0	ND	8.8	ND	ND

[a] Relative improvement in time to pregnancy and absolute improvement in first-cycle pregnancy rate of the ranking method comparisons indicated.
[b] Time to pregnancy measured in number of cycles.
ND, not determined.

cohort ranking method. To carry out this analysis, embryos were randomized to cohorts as described in the Materials and methods section, but no cohorts were excluded. To assess the likelihood of a cohort containing a viable embryo, only the scores of the top-ranked embryo in each cohort were considered. A binary threshold for predicting at least one viable embryo was selected based on the average artificial intelligence score of all top-ranked embryos in the analysis. Predictions of at least one viable embryo were defined as being equal to or higher than the average score, whereas predictions of no viable embryos were defined as being less than the average score.

For the retrospective viability dataset, 2.1% of the approximate 110,000 simulated cohorts contained no viable embryos. The mean score of top-ranked embryos across all cohorts was 9.7/10.0. On the basis of this threshold, the artificial intelligence model could predict if a cohort had at least one viable embryo with a class accuracy of 79.4% (true positive rate), and if a cohort had no viable embryos with a class accuracy of 66.8% (true negative rate). Overall accuracy on this dataset was 79.1%. For the Gardner dataset, 3.8% of the approximately 50,000 cohorts contained no viable embryos, and the mean score of top-ranked embryos was 9.2/10.0. The true positive rate for identifying cohorts with at least one viable embryo was 67.5%, and the true negative rate for identifying cohorts with no viable embryos was 75.0% (overall accuracy of 67.8%). These seemingly higher accuracies than those reported for the artificial intelligence model in predicting viability on a per-embryo basis are explained by the added requirement of selecting an optimal dataset-specific prediction threshold, which would necessitate different clinics optimizing prediction thresholds on their own patient groups over time.

Artificial intelligence scores correlate with embryo ploidy and specific chromosomal abnormalities as determined by PGT-A

Preimplantation genetic testing for aneuploidy is commonly used to evaluate embryo ploidy status, which is known to influence clinical outcomes in IVF (*Scott et al., 2012*; *Dahdouh et al., 2015*; *Schaeffer et al., 2018*). An aim of the present study was to evaluate if an artificial intelligence model that predicts pregnancy outcomes might also detect visible features associated with ploidy status. This was addressed using a retrospective dataset of 5469 day 5 embryo images with matched PGT-A outcomes (Supplementary Table 3). Images were provided by five IVF clinics in the USA for 2615 women who had undergone IVF procedures between 2015 and 2020.

Comparison of artificial intelligence scores and PGT-A outcomes revealed that the average artificial intelligence score was significantly higher for euploid embryos than for aneuploid embryos ($P < 0.0001$), corresponding to a higher potential for successful clinical pregnancy (FIGURE 4A). For mosaic embryos, the average artificial intelligence score fell between the scores for euploid and aneuploid groups, consistent with an intermediate proportion of abnormal cells making up these embryos and a higher potential for successful clinical outcomes over aneuploid embryos owing to the process of self-correction (*Greco et al., 2015*; *Victor et al., 2019*; *Viotti et al., 2020*). Given the high proportion of non-euploid embryos generally identified by PGT-A (approximately 30–60% based on age) (*Viotti, 2020*), experts in the field have started to define specific chromosomal abnormalities that have a lower risk profile for mosaic embryos on transfer, as an alternative option for when no euploid embryos are available for use (*CoGEN, 2017*; *Cram et al., 2019*; *Grati et al., 2018*). Our results showed that the artificial intelligence score was significantly higher ($P = 0.0197$) for embryos that have segmental abnormalities than for those with full chromosomal gains or losses (FIGURE 4B), which is consistent with improved clinical outcomes for these embryos, including increased

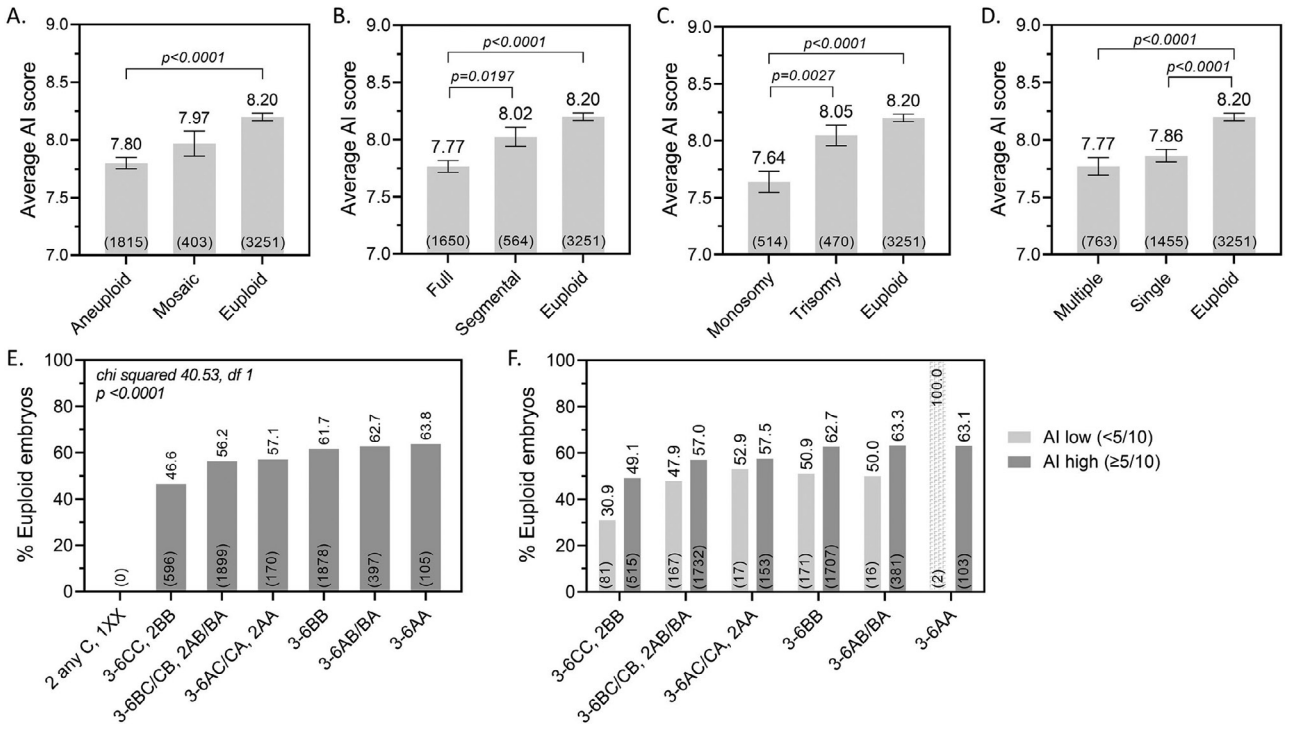

FIGURE 4 Artificial intelligence scores correlate with features of chromosomal abnormality detected by preimplantation genetic testing for aneuploidy (PGT-A). Correlations between average artificial intelligence score and (A) overall euploid, mosaic or aneuploid embryo status; (B) full chromosomal gains or losses, or segmental duplications or deletions ('segmental' subgroup included embryos with single or multiple segmental abnormalities; 'full' subgroup included embryos with full chromosomal abnormalities (gain or loss), either alone or in combination with additional abnormalities); (C) monosomic or trisomic changes (only embryos with single full monosomies or trisomies were included, embryos with multiple abnormalities were excluded); and (D) single or multiple chromosomal abnormalities detected by PGT-A. Average artificial intelligence score is depicted for each group, and the number of embryos in each group is indicated in parentheses (analysis of variance $P < 0.001$ for each comparison); (E) the proportion of euploid embryos according to morphological grade using the Gardner seven-group ranking system. Pregnancy rate is depicted for each group, and the number of embryos in each group indicated in parentheses. Note that no embryos of expansion grade 1 or 2 were available in the PGT-A test dataset; (F) the proportion of euploid embryos in each morphological rank group (seven-group system) according to high (<5/10) or low (≥5/10) artificial intelligence score. The pregnancy rate is depicted for each group, and the number of embryos in each group indicated in parentheses. AI, artificial intelligence, df, degrees of freedom.

implantation and clinical pregnancy rates (*Victor et al., 2019*; *Viotti et al., 2020*). Average artificial intelligence scores were also significantly higher for embryos with trisomic over monosomic changes ($P = 0.0027$), which is consistent with monosomic abnormalities generally being non-viable (*PGDIS, 2016*). Although artificial intelligence scores were not significantly different between embryos with single versus multiple chromosomal abnormalities, a trend was observed towards higher scores for those with single abnormalities (FIGURE 4D). This aligns with recommendations for selecting mosaic embryos with certain single abnormalities for transfer over those with multiple or complex aberrations (*CoGEN, 2017*; *Cram et al., 2019*). Similar correlations were observed when evaluating the proportion of specific abnormalities according to artificial intelligence score brackets (viability likelihood categories) (Supplementary Figure 2).

The proportion of euploid embryos was found to significantly correlate with increasing Gardner-based score groups, demonstrating a relationship between embryo morphology and ploidy status (FIGURE 4E). Interestingly, high artificial intelligence scores were associated with a higher percentage of euploid embryos even when considering embryos of similar morphological grades (FIGURE 4F), suggesting that the artificial intelligence model could potentially be used to differentiate embryos of similar morphology based on ploidy status.

Regions of relevance to artificial intelligence analysis demonstrate some overlap with known morphological features of embryo quality

The observed correlations of artificial intelligence score with morphological components of the embryo and with specific types of chromosomal abnormalities suggest that the artificial intelligence model may be detecting a combination of known and as-yet unknown morphological features. To begin investigating these features, class activation maps (CAM) were generated for an example input image representing an embryo graded 4BA by Gardner score. The embryo received an artificial intelligence score of 9.1 and resulted in clinical pregnancy.

Although a range of CAM algorithms exist, the Grad-CAM++ method was used in this preliminary analysis (*Chattopadhay et al., 2018*). Heat maps associated with regions of high or low relevance to the artificial intelligence prediction are presented for each of the constituent convolutional neural network models making up the overall artificial intelligence ensemble model in FIGURE 5. In general, regions of relevance corresponding to viable and non-viable classifications differed from each other, with a localized focus on the inner cell mass for non-viability and a

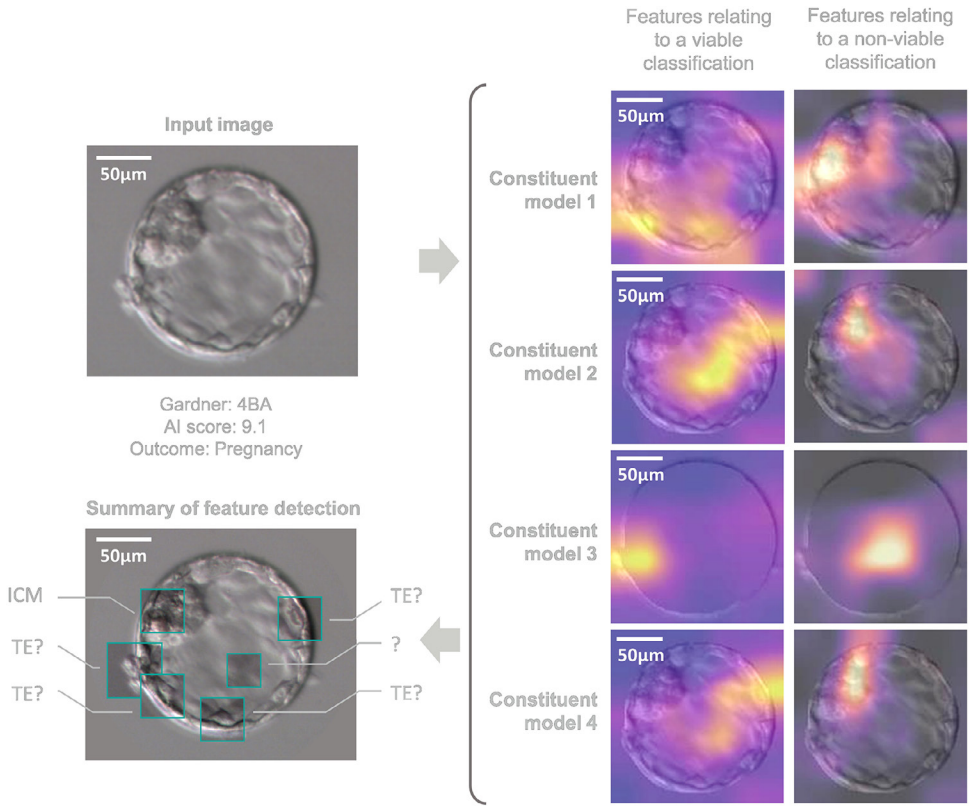

FIGURE 5 Grad-CAM++ mapping of artificial intelligence feature detection for a single day 5 blastocyst image. Heat maps associated with four constituent convolutional neural network models making up the final artificial intelligence model are depicted for regions relating to a viable (left row) and non-viable (right row) classification. The intensity of the regions of relevance is indicated on a scale from yellow to dark purple (for viability), and from white through to grey (for non-viability), corresponding to regions of the most to least relevance during artificial intelligence analysis. The original input image is provided, as is a depiction of the summary of features detected by overlaying heat maps from all constituent models. Regions of focus corresponding to the inner cell mass (ICM) and trophectoderm (TE) are indicated, as are regions where the significance of the detected feature is unclear (indicated with a question mark). The significance of localized focus on specific parts of the trophectoderm is unclear. Note that regions of relevance lying outside the embryo boundary may provide information regarding overall embryo shape (expansion grade). AI, artificial intelligence.

more distributed focus on parts of the trophectoderm and the blastocoele for viability. For the trophectoderm, it is not possible at this stage to determine if the regions of relevance reflect the trophectoderm grade in general, or specific localized parts of the trophectoderm, zona pellucida, or both, for which the significance is unknown, or whether it simply reflects detection of the overall shape of the embryo (expansion grade).

It is important to note that these observations are preliminary, and further investigation of feature detection will be required using multiple mapping methods, different experimental parameters and a range of embryo sub-types to understand more fully what the artificial intelligence model is detecting. The collective results presented in this body of work, however, provide support for the notion that the artificial intelligence model is likely detecting some well-known morphological features, as well as features beyond those that are visually apparent using the Gardner method, which may in some instances correspond to features reflecting chromosomal integrity.

DISCUSSION

Evaluating embryo selection techniques and establishing comparative performance is challenging, because the ground truth pregnancy outcome for each embryo in a patient cohort is not generally known. This has led the field to turn to simple binary comparisons for performance evaluation of artificial intelligence, which do not accurately reflect real-world clinical practice. Binary accuracy measures, such as sensitivity, specificity, precision and recall provide some idea of the likelihood of clinical pregnancy on a per-embryo basis (answering the question, is this embryo likely or unlikely to lead to pregnancy?); however, they do not provide an estimation of the ability of an artificial intelligence model to rank multiple embryos within a patient cohort and select the best one for transfer (answering the question, which embryo is the most likely to lead to pregnancy?). Even ROC-AUC only provides, at best, an approximate estimation of the ranking ability of a binary classification system for this particular application, as the curves generated are simply an integration of the accuracy of the model over the complete range of possible binary threshold values. Therefore, although ROC-AUC is a useful tool for comparing artificial intelligence model performance on any given dataset, it still lacks the ability to estimate more clinically relevant performance measurements like the number of cycles needed to achieve pregnancy, or pregnancy rate. While

ultimately a randomized controlled trial remains the standard paradigm for evaluating interventional methods in medicine, it is not always practicable to conduct such burdensome studies, particularly for evaluating low-risk clinical decision support artificial intelligence, which may undergo future iterations resulting in the need for repeated performance evaluation.

The present study identifies several alternate methods for evaluating artificial intelligence models for ranking embryos to better mirror the process of embryo assessment and selection in the IVF laboratory, without relying on an interventional clinical trial. Correlative analyses linking increasing artificial intelligence score to increasing clinical pregnancy rates demonstrated that the viability artificial intelligence model can be used to aid in ranking embryos within a patient cohort according to relative likelihood of pregnancy. These methods were derived from procedures for developing calibration curves traditionally used to validate the performance of in vitro diagnostic devices.

A simulated cohort analysis was also developed to allow an estimation of TTP using tens of thousands of simulated patient IVF cohorts (cycles). This method showed that the use of the artificial intelligence model improved ranking and reduced the average number of cycles that might be needed to achieve clinical pregnancy. Given that embryo cryopreservation techniques have been optimized over the years such that frozen embryo transfers now demonstrate equivalent or even higher success rates than on fresh transfer, improved embryo selection is unlikely to affect cumulative pregnancy rates in IVF (*Mastenbroek et al., 2011*). As an alternative metric, the simulated cohort ranking method was used to demonstrate that use of the artificial intelligence model improved pregnancy rate in the first cycle of IVF. These improvements were compared with random ranking, which was selected to reflect the approximate 50% pregnancy success rate in the datasets collected for this study, and with a number of different Gardner-based ranking methods, a common grading system typically used to select embryos based on pre-defined morphological characteristics. Regardless of the comparator ranking method used, the artificial intelligence model was found to be superior in reducing TTP and increasing pregnancy rates. Collectively, these methods allowed evaluation of artificial intelligence performance for embryo ranking and selection on a per-patient basis, rather than on a per-embryo basis, which more closely reflects the intended clinical use of the artificial intelligence model for ranking embryos in preferred order for transfer.

The artificial intelligence model also improved embryo selection based on live birth outcome compared with random selection using the simulated cohort ranking method. The improvement was, however, less than when selecting embryos based on a clinical pregnancy outcome. This is not surprising, as there are likely additional confounding factors outside of embryo quality that could contribute to miscarriage after an initial positive clinical pregnancy result, e.g. hormonal conditions, uterine or cervical factors, or infection. A clinical end point of pregnancy is in this case a more appropriate ground truth outcome than live birth for evaluating the performance of artificial intelligence models trained to assess embryo quality.

Artificial intelligence approaches for embryo selection are considered by many to be a black box evaluation. It is important to characterize these models for consistency with known clinical features of embryo quality to gain trust in the predictions they make. This study showed not only correlation of artificial intelligence scores with known morphological features typically associated with embryo viability, but also correlation with embryo ploidy status, providing confidence that the artificial intelligence model is identifying complex features that relate to known metrics of embryo quality that are relevant to pregnancy outcomes. A preliminary investigation of feature detection using CAM methodology supported the hypothesis that the artificial intelligence model likely detects both known and previously unknown morphological features related to embryo quality, although substantial additional work will be required to characterize the regions of relevance and their biological significance more fully. Given the non-linearity of the Gardner grade, and the inconsistency of its application between clinics and embryologists (*Storr et al., 2017*), the results of this study show that the artificial intelligence model offers a more objective and accurate measurement of embryo quality with regards to pregnancy likelihood, which can be used to effectively rank order embryos for transfer. Considering the state of the IVF field today, even relatively small, incremental improvements in success rates are important (*Chambers et al., 2021*). That the artificial intelligence model reduced TTP by up to 12.2% and improved first-cycle pregnancy success rate by up to 11.1% over morphological methods is likely to translate to a significant real-world benefit.

In conclusion, the findings of this study outline methods that should be considered more widely for evaluating artificial intelligence models in the field of embryology, and further support the use of this viability artificial intelligence for embryo selection during IVF treatment.

ACKNOWLEDGEMENTS

The authors acknowledge the kind support of investigators and collaborating clinics for providing embryo images and associated data as follows: Matthew 'Tex' VerMilyea, Andrew Miller, and Roberta Hanson, Ovation Fertility (Austin TX and San Antonio TX, USA); Glen Adaniya, RaeAnne van Tol, and Bradford Bopp, Midwest Fertility Specialists (Carmel IN, USA); Rebecca Matthews and Brandon Bankowski, ORM Fertility (Portland OR, USA); Erica Behnke, Institute for Reproductive Health (Cincinnati OH, USA); Joan Riley, Washington University (St Louis MO, USA); Adelle Yun Xin Lim, Alpha IVF & Women's Specialists (Petaling Jaya, Malaysia); Wiwat Quangkananurug and Sujin Chanchamroen, Safe Fertility (Bangkok, Thailand); Jon Aizpurua and Lydia Giardini, IVF-Spain (Alicante, Spain); Kelli Sorby, No. 1 Fertility (Melbourne VIC, Australia); Dean Morbeck, Fertility Associates (Auckland, Christchurch, Dunedin, Hamilton, and Wellington, New Zealand); Ashleigh Storr, Flinders Fertility (Adelaide SA, Australia); and Hamish Hamilton and Michelle Lane, Repromed (Adelaide SA and Tiwi NT, Australia).

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rbmo.2022.07.018.

REFERENCES

Berntsen, J., Rimestad, J., Lassen, J.T., Tran, D., Kragh, M.F. **Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences.** PLoS One 2022; 17e0262661

Capalbo, A., Rienzi, L., Cimadomo, D., Maggiulli, R., Elliott, T., Wright, G., Nagy, Z.P., Ubaldi, F.M. **Correlation between standard blastocyst morphology, euploidy and implantation: an observational study in two centers involving 956 screened blastocysts.** Hum. Reprod. 2014; 29: 1173–1181

Chambers, G.M., Dyer, S., Zegers-Hochschild, F., de Mouzon, J., Ishihara, O., Banker, M., Mansour, R., Kupka, M.S., Adamson, G.D. **International Committee for Monitoring Assisted Reproductive Technologies world report: assisted reproductive technology, 2014†.** Hum. Reprod. 2021; 36: 2921–2934

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N. **Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks.** 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) 12-15 March 2018: 839–847

Chavez-Badiola, A., Flores-Saiffe-Farías, A., Mendizabal-Ruiz, G., Drakeley, A.J., Cohen, J. **Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation.** Reproductive BioMedicine Online 2020; 41: 585–593

CoGEN. COGEN position statement on chromosomal mosaicism detected in preimplantation blastocyst biopsies. 2017 [15 Jan 2021]. Available from: https://www.ivf-worldwide.com/index.php?option=com_content&view=article&id=733&Itemid=464.

Cram, D.S., Leigh, D., Handyside, A., Rechitsky, L., Xu, K., Harton, G., Grifo, J., Rubio, C., Fragouli, E., Kahraman, S., Forman, E., Katz-Jaffe, M., Tempest, H., Thornhill, A., Strom, C., Escudero, T., Qiao, J., Munne, S., Simpson, J.L., Kuliev, A. **PGDIS Position Statement on the Transfer of Mosaic Embryos 2019.** Reprod. Biomed. Online 2019; 39: e1–e4

Dahdouh, E.M., Balayla, J., García-Velasco, J.A. **Comprehensive chromosome screening improves embryo selection: a meta-analysis.** Fertil. Steril. 2015; 104: 1503–1512

Erlich, I., Ben-Meir, A., Har-Vardi, I., Grifo, J., Wang, F., McCaffrey, C., McCulloh, D., Or, Y., Wolf, L. **Pseudo contrastive labeling for predicting IVF embryo developmental potential.** Sci. Rep. 2022; 12: 2488

Florkowski, C.M. **Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests.** Clin. Biochem. Rev. 2008; 29 Suppl 1: S83–S87

Gardner, D.K., Schoolcraft, W. **In Vitro Culture of Human Blastocyst.** Jansen R., Mortimer D. Towards Reproductive Certainty: Infertility and Genetics Beyond Parthenon Press Carnforth 1999: 377–388

Gardner, D.K., Lane, M., Stevens, J., Schlenker, T., Schoolcraft, W.B. **Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer.** Fertil. Steril. 2000; 73: 1155–1158

Gardner, D.K., Stevens, J., Sheehan, C.B., Schoolcraft, W. **Analysis of blastocyst morphology.** Elder KCJ Human Preimplantation Embryo Selection Informa Healthcare London 2007: 79–87

Gardner, D.K., Balaban, B. **Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'OMICS': is looking good still important?** MHR: Basic science of reproductive medicine 2016; 22: 704–718

Grati, F.R., Gallazzi, G., Branca, L., Maggi, F., Simoni, G., Yaron, Y. **An evidence-based scoring system for prioritizing mosaic aneuploid embryos following preimplantation genetic screening.** Reprod. Biomed. Online 2018; 36: 442–449

Greco, E., Minasi, M.G., Fiorentino, F. **Healthy Babies after Intrauterine Transfer of Mosaic Aneuploid Blastocysts.** N. Engl. J. Med. 2015; 373: 2089–2090

Hara, T., Katsuki, T., Kusuda, T., Ohama, K. **Pregnancy rate, multiple pregnancy rate, and embryo quality: Clues for single blastocyst transfer from double blastocyst transfer in an unselected population.** Reprod. Med. Biol. 2005; 4: 153–160

Irani, M., Reichman, D., Robles, A., Melnick, A., Davis, O., Zaninovic, N., Xu, K., Rosenwaks, Z. **Morphologic grading of euploid blastocysts influences implantation and ongoing pregnancy rates.** Fertil. Steril. 2017; 107: 664–670

Kemper, J.M., Liu, Y., Afnan, M., Hammond, E.R., Morbeck, D.E., Mol, B.W.J. **Should we look for a low-grade threshold for blastocyst transfer? A scoping review.** Reproductive BioMedicine Online 2021; 42: 709–716

Khosravi, P., Kazemi, E., Zhan, Q., Malmsten, J.E., Toschi, M., Zisimopoulos, P., Sigaras, A., Lavery, S., Cooper, L.A.D., 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. 2019.; 2: 21

Kozubowski, T., Podgorski, K. **A Multivariate and Asymmetric Generalization of Laplace Distribution.** Computational Statistics 2000; 15: 531–540

Loewke, K., Cho, J.H., Brumar, C.D., Maeder-York, P., Barash, O., Malmsten, J.E., Zaninovic, N., Sakkas, D., Miller, K.A., Levy, M., VerMilyea, M.D. **Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos.** Fertil. Steril. 2022; 117: 528–535

Mastenbroek, S., van der Veen, F., Aflatoonian, A., Shapiro, B., Bossuyt, P., Repping, S. **Embryo selection in IVF.** Hum. Reprod. 2011; 26: 964–966

PGDIS. PGDIS Position statement on chromosome mosaicism and preimplantation aneuploidy testing at the blastocyst stage 2016 [15 Jan 2021]. Available from: https://www.pgdis.org/docs/newsletter_071816.html.

Schaeffer, E., Porchia, L., Lopez-Luna, A., Hernandez-Melchor, D., Lopez-Bayghen, E. **Aneuploidy rates inversely correlate with implantation during in vitro fertilization procedures: In favor of PGT.** Gomy I Modern Medical Genetics and Genomics IntechOpen 2018

Scott, R.T. Jr., Ferry, K., Su, J., Tao, X., Scott, K., Treff, N.R. **Comprehensive chromosome screening is highly predictive of the reproductive potential of human embryos: a prospective, blinded, nonselection study.** Fertil. Steril. 2012; 97: 870–875

Silver, D.H., Feder, M., Gold-Zamir, Y., Polsky, A.L., Rosentraub, S., Shachor, E., Weinberger, A., Mazur, P., Zukin, D., Bronstein, A.M. **Data-Driven Prediction of Embryo Implantation Probability Using IVF Time-lapse Imaging.** ArXiv 2020; 2006.01035 https://arxiv.org/abs/2006.01035

Storr, A., Venetis, C.A., 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: 307–314

Tran, D., Cooke, S., Illingworth, P.J., Gardner, D.K. **Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer.** Hum. Reprod. 2019; 34: 1011–1018

VerMilyea, M., Hall, J.M.M., Diakiw, S.M., Johnston, A., Nguyen, T., Perugini, D., Miller, A., Picou, A., Murphy, A.P., 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: 770–784

Victor, A.R., Tyndall, J.C., Brake, A.J., Lepkowsky, L.T., Murphy, A.E., Griffin, D.K., McCoy, R.C., Barnes, F.L., Zouves, C.G., Viotti, M. **One hundred mosaic embryos transferred prospectively in a single clinic: exploring when and why they result in healthy pregnancies.** Fertil. Steril. 2019; 111: 280–293

Viotti, M., Victor, A.R., Barnes, F.L., Minasi, M.G., Greco, E., Munné, S. **New insights from one thousand mosaic embryo transfers: Features of mosaicism dictating rates of implantation, spontaneous abortion, and neonate health.** Fertil. Steril. 2020; 114: E1–E2

Viotti, M. **Preimplantation Genetic Testing for Chromosomal Abnormalities: Aneuploidy, Mosaicism, and Structural Rearrangements.** Genes (Basel) 2020; 11

Zhao, Y.Y., Yu, Y., Zhang, X.W. **Overall Blastocyst Quality, Trophectoderm Grade, and Inner Cell Mass Grade Predict Pregnancy Outcome in Euploid Blastocyst Transfer Cycles.** Chin. Med. J. (Engl.) 2018; 131: 1261–1267

Received 2 May 2022; received in revised form 30 June 2022; accepted 25 July 2022.