Artificial Intelligence (AI) technology can predict human embryo viability across multiple laboratories with varying demographics with high accuracy and reproducibility

Matthew D. VerMilyea, et al.

ESHRE 2019 Vienna, Austria

Artificial Intelligence (AI) technology can predict human embryo viability across multiple laboratories with varying demographics with high accuracy and reproducibility

O-004 Matthew D. VerMilyea, et al. ESHRE 2019 Vienna, Austria

INTRODUCTION

Over the past 41 years, success rates with IVF have drastically improved. However, the ability to accurately and cost-effectively de-select embryos with a lower success potential from a cohort remains a challenge. Advances in genetic competency evaluation by pre-implantation genetic testing (PGT) and other non-invasive technologies, such as time-lapse morphokinetic assessment and spent media analysis, have proven to be cost prohibitive for some while the accuracy of such assessments remain debatable[1,2,3]. Clinicians and embryologists continue to default to traditional morphological assessment and embryo grading protocols which have been used since the inception of IVF[4]. Using a standard phase contrast microscope, developmental milestones of the embryo are evaluated and assessments regarding embryo viability potential are subjectively made by skilled embryologists. As part of their embryo assessment process, it has also become routine for embryologists to obtain a static digital 2D image of a blastocyst prior to transfer, biopsy or vitrification.

Our study is based on the technical analysis of the Life Whisperer AI application and its ability to predict Day 5 blastocyst viability. Viability for our study was measured by clinical pregnancy, as defined by fetal heartbeat at first ultrasound. Embryo images and data were collected across multiple clinics and geographical sites in North America, Australia, Malaysia, and New Zealand.

This study is composed of two phases. The purpose of phase one was to demonstrate suitability of the AI method in predicting embryo viability. Phase two of the study represents a multi-center investigation assessing the AI across several clinics in different geographical locations to demonstrate the transferability and generalizability of the AI approach.

The overall study design involved collection of 8,948 Day 5 blastocysts images. 8,886 images were obtained from phase contrast microscopes equipped with a standard camera. This type of equipment is traditional of most IVF laboratories and inexpensive to procure and maintain. All embryo images were collected prior to freezing or biopsy for PGT analysis and were of a minimum resolution of 512 x 512 pixels with the entire embryo in the field of view.

Data were obtained for consecutive patients who had undergone IVF at 12 independent clinics from 2011 to 2018. Data were limited to patients who received a single embryo transfer with a Day 5 blastocyst, and where the endpoint was clinical pregnancy outcome. The clinical pregnancy endpoint was deemed to be the measure most reliant on the viability of the embryo with limited confounding patient related factors post-implantation. Of note, anecdotal evidence from clinics suggest that approximately 20% of an unsuccessful IVF outcome is due to factors unrelated to the viability of the embryo[5]. These factors include operational errors, and patient related factors including, but not limited to, uterine pathologies and endometriosis. As a result, the theoretical maximum accuracy of predictions made by the AI is considered to be 80%. For a subset of patients, the embryologist morphokinetic grade was known, and was used to compare the accuracy of the AI with the standard visual grading method for those patients.

All patient data provided by clinics for this study was de-identified and the study was deemed exempt from IRB review pursuant to the terms of the U.S. Department of Health and Human Service's Policy for Protection of Human Research Subjects at 45 C.F.R. 46.101(b).

RESULTS

Phase One Design

Phase one was performed using a total of 4650 Day 5 blastocyst images. The purpose of this phase was to demonstrate suitability of the AI method in predicting embryo viability. When training an AI machine-learning-based model, data need to be split into three separate groups. These groups consist of a training dataset, validation dataset, and blind test dataset. Training data are used to train the AI model, the

validation data inform the selection of the best AI model with the highest level of accuracy, and the blind data comprise a completely independent data set that has not been used either in the training of the model or in the selection of the best model. Reporting on this blinded set is the true measure of the AI accuracy as it represents a set of images that has been held back prior to AI training, and therefore the AI model has never analyzed nor does the model know the outcome. Random assignment was used to split the data among the dataset categories, ensuring that the ratio of viable to non-viable examples is uniform for each dataset category. Two separate blind test datasets were used for testing in this study. Not all embryo images had recorded embryologist grading score information, as shown in Table 1. Table 1 also provides a breakdown of the data available and used in each of the dataset categories.

	Training Dataset	Validation Dataset	Blind Test Set 1	Blind Test Set 2
Number of images	3892 (80% of total)	390 (10% of total)	368 (9.5% of total)	632 (Additional)
Number of positive clinical pregnancies		70 (17.9%)	76 (20.7%)	194 (30.7%)
Number of negative clinical pregnancies		320 (82.1%)	292 (79.3%)	438 (69.3%)
Number of images with embryologist grade		149	121	477

Table 1. AI training, validation and test dataset descriptions.

The AI deep learning model was trained on the Training Dataset. The trained AI was then applied to the validation set in order to fine-tune the model during training, and therefore represents a biased set used as part of the training procedure. The model is then applied to the two blinded test datasets (Blind Test Set 1 and 2) to assess predictive accuracy. Accuracy in identifying viable embryos (sensitivity) is calculated as the number of embryos that the AI identified as viable divided by the total number of known viable embryos that resulted in a positive clinical pregnancy. Similarly, accuracy in identification of non-viable embryos (specificity) is calculated as the number of embryos that the AI identified as non-viable divided by the total number of known non-viable embryos that resulted in a negative clinical pregnancy outcome. An AI viability score above 50% is considered viable, and equal to or below 50% is considered non-viable.

Overall Accuracy by Sensitivity and Specificity

The most important measure of accuracy is that reported for the Blind Test Sets. The average AI model sensitivity across the two Blind Test Sets is 74.1% and the specificity is 65.3% (**Figure 1**). The accuracy values are considered high given the overall positive pregnancy rate in this dataset was 27% across all transferred embryos that had been pre-selected by embryologists.

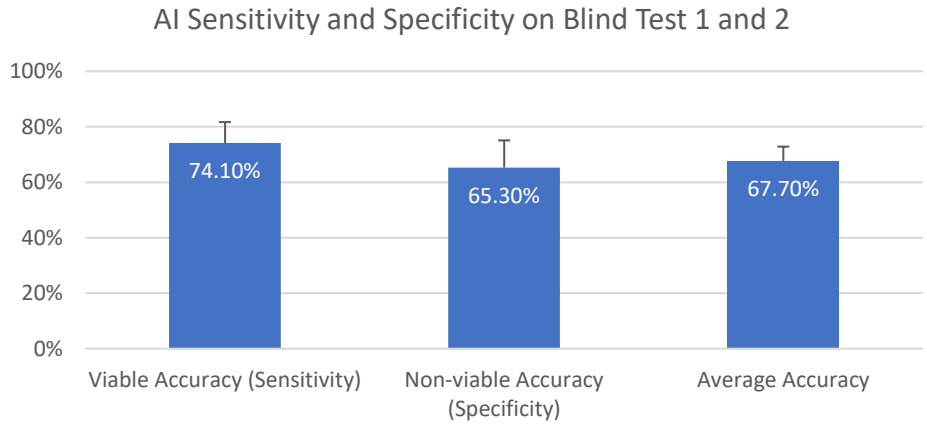

Figure 1. AI model sensitivity and specificity on Blind Test Set 1, Blind Test Set 2, and the average of the Blind Test Sets.

AI Versus Embryologist Accuracy Comparison

In addition to assessing the accuracy of the AI on predicting embryo viability in the individual test sets, a comparison was conducted to test the AI accuracy compared to the embryologist's grade (**Figure 2**). This provides a measure of accuracy of AI versus accuracy of the embryologist in scoring embryos as viable or non-viable on the basis of standard morphokinetic grading.

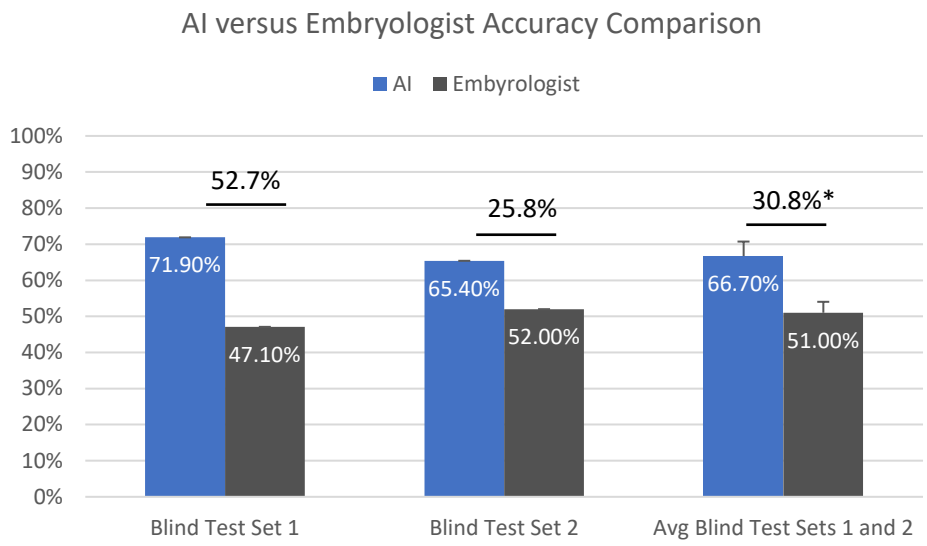

Figure 2. Accuracy of AI compared with embryologist morphokinetic grading for identifying embryo viability. *$p=0.04$, $n=2$.

Phase Two Design

The purpose of phase two was to demonstrate the generalizability of the AI to different clinical environments and equipment in different geographical locations.

Phase two included a total of 3604 Day 5 blastocyst images from 12 clinics in USA, Australia, and New Zealand. Of these images, 2217 were separated to form three subsets in the same manner as in the first study: The Training Dataset, Validation Dataset and Blind Test Set 1.

	Training Dataset	Validation Dataset	Blind Test Set 1	Blind Test Set 2	Blind Test Set 3
Number of images	1744 (79% of total)	193 (9% of total)	280 (13% of total)	286	1101
Number of positive clinical pregnancies	858 (49.2%)	97 (50.3%)	141 (50.4%)	180	334
Number of negative clinical pregnancies	886 (50.8%)	96 (49.7%)	139 (49.6%)	106	767
Number of images with embryologist grade	1529	177	262	0	539
Average patient age	33.9	33.6	34.2	N/A	34.4

Table 2. AI training and test dataset descriptions.

Overall Accuracy

A summary of the total accuracy can be found in **Table 3** for Life Whisperer Model v1.0, as applied to the mixed demographic Blind Test Set 1 comprising data from multiple clinics. The overall accuracy was consistent with that of phase one indicating the AI is transferable to a wider more diverse clinical dataset.

Blind Test Set 1	Viable	Non-Viable	Total
Model accuracy	102/141 = **72.34%**	76/139 = **54.68%**	178/280 = **63.57%**

Table 3. Accuracy of the Life Whisperer Model v1.0, when applied to the Blind Test Set 1. Results show the accuracy in identifying viable embryos, non-viable embryos, and the total accuracy for both viable and non-viable embryos combined.

Detailed AI Assessment – Individual Clinic Analysis

Although the overall accuracy of the AI is useful for assessing the efficacy of the model, we also wanted to analyze the results when various clinic groups were broken down. Results showed a high level of accuracy when the AI was applied to individual clinic groups demonstrating the ability of the AI to translate across diverse clinical environments and different patient demographics.

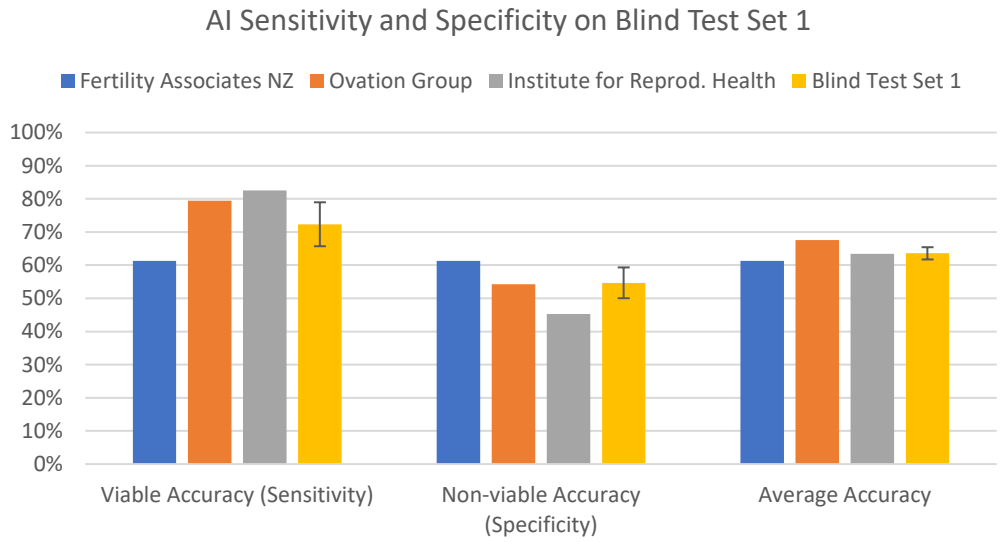

Figure 3. Accuracy of AI compared across the individual clinics involved in Blind Test Set 1. Final column within each grouping represents the total combined set.

DISCUSSION AND OVERALL CONCLUSIONS

A two-phase study was conducted to assess the relative accuracy improvement provided by the Life Whisperer AI embryo assessment tool compared to standard morphokinetic grading by embryologists. For the first phase, Life Whisperer's AI was trained on 3892 images. The validation and testing of the AI was then assessed using two blind datasets from a total of 1000 embryo images that had not been used during the AI training process. Results showed the AI had an overall accuracy of 67.7% in identifying embryo viability across the two blind datasets. Accuracy was defined as the sum of the number of embryos that the AI blindly predicted were viable and that had a positive pregnancy outcome (sensitivity), in addition to the number of embryos that the AI blindly predicted were non-viable and that resulted in a negative pregnancy outcome (specificity), divided by the total number of embryos. The sensitivity of the AI for viable embryos was 74.1% and the specificity was 65.3%. The accuracy improvement of the AI compared with the embryologist grading method alone was calculated using the overall accuracy of the embryologist grading (51%) compared to that of the AI (67.7%). The accuracy improvement demonstrated by the AI was therefore 30.8% compared with embryologist grading.

For the second phase, 3604 embryo images from multiple demographics were assessed. Of these images, 2,111 had associated embryologist's grades available. This study involved training the AI on 1744 independent embryo images and clinical pregnancy outcomes, and then blind testing on data from the individual IVF clinics. The purpose of this study phase was to assess the transferability of the AI to different clinical environments and demographics.

Results from phase two showed a high level of transferability of the AI to different clinical environments and the ability to maintain >25% accuracy improvement across different clinical settings compared with embryologist grading. The Blind set showed the AI sensitivity of 72.3% and specificity of 54.7%.

Overall, the AI approach tested showed high specificity and sensitivity for assessment of Day 5 embryos and predictability of clinical pregnancy outcomes. The AI approach showed an average 29.9% accuracy improvement over traditional manual and visual grading methods currently used by embryologists. The value of the AI approach is not only in accuracy improvement of embryo viability assessment, but also in the standardization of embryo assessment where manual embryologist grading is highly subjective with high levels of intra- and inter- operator variability.

This comprehensive analysis in multiple clinical environments, spanning various microscope and camera equipment, and in different demographic locations, provides strong evidence that Life Whisperer AI delivers an improved predictive accuracy in classifying embryo viability. The implication of these results is that the AI can be used to inform selection of embryos, suggesting that this improved accuracy in selection of the best embryo for a given patient will result in improved pregnancy success rates and reduce the overall number of cycles leading to a pregnancy for a given patient.

REFERENCES

1. Gleicher, N., Kushnir, V.A. and Barad, D.H., 2018. How PGS/PGT-A laboratories succeeded in losing all credibility. *Reproductive biomedicine online*, *37*(2), pp.242-245.
2. Racowsky, C., Kovacs, P. and Martins, W.P., 2015. A critical appraisal of time-lapse imaging for embryo selection: where are we and where do we need to go?. *Journal of assisted reproduction and genetics*, *32*(7), pp.1025-1030.
3. Belandres, D., Shamonki, M. and Arrach, N., 2019. Current status of spent embryo media research for preimplantation genetic testing. *Journal of assisted reproduction and genetics*, pp.1-8.
4. Edwards, R.G., Purdy, J.M., Steptoe, P.C. and Walters, D.E., 1981. The growth of human preimplantation embryos in vitro. *American Journal of Obstetrics and Gynecology*, *141*(4), pp.408-416.
5. Annan, J.J.K., Gudi, A., Bhide, P., Shah, A. and Homburg, R., 2013. Biochemical pregnancy during assisted conception: a little bit pregnant. *Journal of clinical medicine research*, *5*(4), p.269.