CLINICAL VALIDATION OF AN ARTIFICIAL INTELLIGENCE (AI) MODEL FOR EMBRYO EVALUATION ACROSS TWO TIME-LAPSE IMAGING SYSTEMS

Authors: Curchoe, CL (1, 2), Shapiro M (1), Gilboa, D (1), Tauber Y (1), Seidman DS (1)

Affiliations: (1) AIVF, Tel Aviv, Israel
(2) ART Compass, Newport Beach, CA

Background

Manually grading embryos is considered the gold standard for embryo evaluation. This method, however, is time-consuming and fraught with observer variability. The AIVF Day-5 embryo evaluation tool, a component of the EMA™ AI-powered operating system by AIVF, provides its users with quantitative assessments of embryo quality. The model's architecture is a convolutional neural network (CNN) ResNet50 backbone with time specialized kinetics heuristics. Clinical validation of the tool is a necessary component of its evaluation strategy to confirm robust performance across various conditions. Here, we assess AIVF Day-5 performance across two time-lapse imaging systems: EmbryoScope+™ (Vitrolife, Denmark) and Geri® (Genea Biomedx, Australia) to validate its technical flexibility and clinical utility.

Objective

To validate the performance of the AIVF Day-5 model across two time-lapse technology systems inside the same IVF clinic. The model outputs a scalar (1 to 9.9) rating for every embryo in the cohort that correlates with its quality and likelihood of clinical pregnancy.

Materials and Methods

Two AIVF Day-5 models with identical architectures were trained and tested on historical embryo images derived from either the Embroscope+ or Geri system, respectively. Two independent datasets with matched characteristics and metadata derived from one European clinic operating both imaging systems were used to assess and compare model performance (Embryoscope, N= 814 embryos; Geri, N=908 embryos). The distribution of scores were compared to conventional morphology grades, ploidy status, and clinical outcomes.

Results

When tested on the independent Geri dataset, there was a significant linear correlation between scores and annotated embryo qualities (graded A-D; ASEBIR), as well as strong ability to differentiate between discrete grades with high classification accuracy (>89%) and significance ($p<0.02$). Average scores were consistently higher for euploid embryos than for aneuploid embryos (reported as mean ± SD: aneuploid: 5.0 ± 0.7; euploid: 5.7 ± 0.3) ($p<0.05$). To assess the association between scores and clinical outcome, the proportion of embryos resulting in clinical pregnancy relative to defined score quartiles ('score bins') was analyzed; a significant linear correlation was observed between scores and the percentage of embryos that resulted in pregnancy (fetal heartbeat) ($p<0.05$). The same analyses were performed on the independent Embryoscope dataset. Scores correlated linearly with the conventional quality scales for all three predefined categories of embryo morphology: integrity of inner cell mass [ICM] and trophectoderm [TE] cells, degree of blastocyst expansion) ($p<0.05$). Average scores were consistently higher for euploid embryos than for aneuploid embryos (aneuploid: 3.7 ± 1.8; euploid: 4.9 ± 1.6) ($p<0.05$); a significant correlation was found between scores and the proportion of embryos that result in pregnancy ($p<0.05$).

Conclusions

AIVF Day-5 for embryo evaluation displays consistent flexibility across two tested imaging systems without compromising on clinical outcomes or utility. The model's strong performance highlights its ability to apply its learned features objectively and robustly to optimize decision-making inside the IVF clinic, regardless of the time-lapse system used.

Financial Support

Curchoe, CL, Shapiro M, Gilboa, D, Tauber Y, Seidman DS, are employers of AIVF.

References: N/A