Human Reproduction, Vol.37, No.10, pp. 2275–2290, 2022
Advance Access Publication on August 9, 2022 https://doi.org/10.1093/humrep/deac171

human reproduction

ORIGINAL ARTICLE *Embryology*

Embryologist agreement when assessing blastocyst implantation probability: is data-driven prediction the solution to embryo assessment subjectivity?

Daniel E. Fordham [1,*,†], Dror Rosentraub [1,†], Avital L. Polsky [1], Talia Aviram [1], Yotam Wolf [1], Oriel Perl [1], Asnat Devir [1], Shahar Rosentraub [1], David H. Silver [1], Yael Gold Zamir [1], Alex M. Bronstein [1,2], Miguel Lara Lara [3], Jara Ben Nagi [4], Adrian Alvarez [5], and Santiago Munné [5]

[1]Embryonics, Embryonics R&D Center, Haifa, Israel [2]Computer Science, Technion—Israel Institute of Technology, Haifa, Israel [3]Unidad de Reproducción Centro Gutenberg, Málaga, Spain [4]Centre for Reproductive and Genetic Health, London, UK [5]Overture Life, AI Department, New York, NY, USA

*Correspondence address. Embryonics, Embryonics R&D Center, Sirkin St. 21, Haifa 3311508, Israel. E-mail: dan@embryonics.me
https://orcid.org/0000-0002-5394-7973

Submitted on March 23, 2022; resubmitted on June 21, 2022; editorial decision on June 30, 2022

STUDY QUESTION: What is the accuracy and agreement of embryologists when assessing the implantation probability of blastocysts using time-lapse imaging (TLI), and can it be improved with a data-driven algorithm?

SUMMARY ANSWER: The overall interobserver agreement of a large panel of embryologists was moderate and prediction accuracy was modest, while the purpose-built artificial intelligence model generally resulted in higher performance metrics.

WHAT IS KNOWN ALREADY: Previous studies have demonstrated significant interobserver variability amongst embryologists when assessing embryo quality. However, data concerning embryologists' ability to predict implantation probability using TLI is still lacking. Emerging technologies based on data-driven tools have shown great promise for improving embryo selection and predicting clinical outcomes.

STUDY DESIGN, SIZE, DURATION: TLI video files of 136 embryos with known implantation data were retrospectively collected from two clinical sites between 2018 and 2019 for the performance assessment of 36 embryologists and comparison with a deep neural network (DNN).

PARTICIPANTS/MATERIALS, SETTING, METHODS: We recruited 39 embryologists from 13 different countries. All participants were blinded to clinical outcomes. A total of 136 TLI videos of embryos that reached the blastocyst stage were used for this experiment. Each embryo's likelihood of successfully implanting was assessed by 36 embryologists, providing implantation probability grades (IPGs) from 1 to 5, where 1 indicates a very low likelihood of implantation and 5 indicates a very high likelihood. Subsequently, three embryologists with over 5 years of experience provided Gardner scores. All 136 blastocysts were categorized into three quality groups based on their Gardner scores. Embryologist predictions were then converted into predictions of implantation (IPG ≥ 3) and no implantation (IPG ≤ 2). Embryologists' performance and agreement were assessed using Fleiss kappa coefficient. A 10-fold cross-validation DNN was developed to provide IPGs for TLI video files. The model's performance was compared to that of the embryologists.

MAIN RESULTS AND THE ROLE OF CHANCE: Logistic regression was employed for the following confounding variables: country of residence, academic level, embryo scoring system, log years of experience and experience using TLI. None were found to have a statistically significant impact on embryologist performance at $\alpha = 0.05$. The average implantation prediction accuracy for the embryologists was 51.9% for all embryos (N = 136). The average accuracy of the embryologists when assessing top quality and poor quality embryos

†The authors consider that the first two authors should be regarded as joint First Authors.

© The Author(s) 2022. Published by Oxford University Press on behalf of European Society of Human Reproduction and Embryology. All rights reserved.
For permissions, please email: journals.permissions@oup.com

(according to the Gardner score categorizations) was 57.5% and 57.4%, respectively, and 44.6% for fair quality embryos. Overall interobserver agreement was moderate ($\kappa = 0.56$, N = 136). The best agreement was achieved in the poor + top quality group ($\kappa = 0.65$, N = 77), while the agreement in the fair quality group was lower ($\kappa = 0.25$, N = 59). The DNN showed an overall accuracy rate of 62.5%, with accuracies of 62.2%, 61% and 65.6% for the poor, fair and top quality groups, respectively. The AUC for the DNN was higher than that of the embryologists overall (0.70 DNN vs 0.61 embryologists) as well as in all of the Gardner groups (DNN vs embryologists—Poor: 0.69 vs 0.62; Fair: 0.67 vs 0.53; Top: 0.77 vs 0.54).

LIMITATIONS, REASONS FOR CAUTION: Blastocyst assessment was performed using video files acquired from time-lapse incubators, where each video contained data from a single focal plane. Clinical data regarding the underlying cause of infertility and endometrial thickness before the transfer was not available, yet may explain implantation failure and lower accuracy of IPGs. Implantation was defined as the presence of a gestational sac, whereas the detection of fetal heartbeat is a more robust marker of embryo viability. The raw data were anonymized to the extent that it was not possible to quantify the number of unique patients and cycles included in the study, potentially masking the effect of bias from a limited patient pool. Furthermore, the lack of demographic data makes it difficult to draw conclusions on how representative the dataset was of the wider population. Finally, embryologists were required to assess the implantation potential, not embryo quality. Although this is not the traditional approach to embryo evaluation, morphology/morphokinetics as a means of assessing embryo quality is believed to be strongly correlated with viability and, for some methods, implantation potential.

WIDER IMPLICATIONS OF THE FINDINGS: Embryo selection is a key element in IVF success and continues to be a challenge. Improving the predictive ability could assist in optimizing implantation success rates and other clinical outcomes and could minimize the financial and emotional burden on the patient. This study demonstrates moderate agreement rates between embryologists, likely due to the subjective nature of embryo assessment. In particular, we found that average embryologist accuracy and agreement were significantly lower for fair quality embryos when compared with that for top and poor quality embryos. Using data-driven algorithms as an assistive tool may help IVF professionals increase success rates and promote much needed standardization in the IVF clinic. Our results indicate a need for further research regarding technological advancement in this field.

STUDY FUNDING/COMPETING INTEREST(S): Embryonics Ltd is an Israel-based company. Funding for the study was partially provided by the Israeli Innovation Authority, grant #74556.

TRIAL REGISTRATION NUMBER: N/A.

Key words: artificial intelligence / deep learning / embryo quality / embryology / implantation / IVF/ICSI outcome / assisted reproduction / embryologist agreement

Introduction

The successful outcome of IVF treatment can be measured in many ways, but ultimately the goal is a live, healthy birth. Terminal outcomes are most often reported as live birth rate (LBR) or cumulative live birth rate (CLBR) (Maheshwari *et al.*, 2015). Multiple factors are known to affect LBR and CLBR and, as such, 40 years of evidence-based advancements in these areas have led to significant improvements in IVF success. These include advances in the understanding of ovarian reserve and ovarian stimulation (van Wely *et al.*, 2011; Broer *et al.*, 2013; Lambalk *et al.*, 2017), improvements in sperm selection and ICSI (Balaban *et al.*, 2011; Walls *et al.*, 2012; Worrilow *et al.*, 2013), improvements in embryo culture and transfer, including elective single embryo transfer (Gardner *et al.*, 1998; Papanikolaou *et al.*, 2006; Thompson *et al.*, 2013) and advances in embryo grading and selection (Gardner *et al.*, 2000; Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology, 2011; Van den Abbeel *et al.*, 2013).

However, although success rates have continued to improve along with the demand for IVF, there is a general consensus that there is still room for improvement: LBRs per intended oocyte retrieval in the USA in 2019 were ~37% (2019 Assisted Reproductive Technology Fertility Clinic and National Summary Report | CDC, 2022). Commentators in the field have consistently pointed to the subjective and variable nature of the embryo selection process as a particular barrier, despite attempts to standardize and validate this approach (Van den Abbeel *et al.*, 2013; Rocha *et al.*, 2016; VerMilyea *et al.*, 2020). Furthermore, direct measurement of the accuracy of embryo selection is challenging since the clinical outcome of non-transferred embryos cannot be known. Instead, agreement is often used to assess performance relative to variable outcomes (Storr *et al.*, 2017; Hammond *et al.*, 2020). The subjective nature of embryo grading is therefore biased further by the fact that performance is typically validated in the context of a subjective gold standard.

Preimplantation genetic testing for aneuploidy (PGT-A) and its predecessors introduced a valuable new tool to the IVF laboratory. As aneuploidy is believed to be a particularly common cause of miscarriage, and incidence increases with age, PGT-A and selective transfer of euploid embryos have been shown to improve ongoing pregnancy (OP) rate per embryo transfer (Munné *et al.*, 2019). Additionally, when euploid embryos are transferred, implantation rates have been observed to be similar regardless of maternal age at the time of oocyte retrieval (Harton *et al.*, 2013). However, although PGT-A has been widely adopted in the USA, its cost and the technicality of the procedure are significant barriers (Roche *et al.*, 2021). Furthermore, it is still unclear whether its use has a significant impact on OP per intention to treat and LBRs (Munné, 2018; Munné *et al.*, 2019; Roche *et al.*, 2021). The current debate surrounds the fate of mosaic embryos, the transfer of which has been traditionally avoided, and guidance is required in this area (Munné *et al.*, 2017; Roche *et al.*, 2021). Regardless, it is still an invaluable tool and provides an objective approach to embryo selection, particularly in cases of advanced maternal age (Munné *et al.*, 2019).

The introduction of time-lapse imaging (TLI) was heralded as another major step in the right direction for embryo culture, enabling clinicians to maintain a constant environment whilst monitoring the entirety of the embryo's *in vitro* development. Unfortunately, its effect on birth rate is again unclear, with a lack of quality evidence suggesting any positive impact (Kirkegaard *et al.*, 2015; Armstrong *et al.*, 2019). TLI has, however, led to improved understanding of morphokinetics and to the accumulation of vast amounts of visual data of embryogenesis. It is likely that certain irregular morphokinetic events, such as reverse cleavage, direct cleavage and multinucleation, as well as the timing of developmental events, are associated with implantation outcome (Kaser and Racowsky, 2014; Motato *et al.*, 2016; Barrie *et al.*, 2017). A recent multicenter double-blinded randomized control trial assessing a commercially available TLI-based algorithm designed to predict implantation outcomes from morphokinetic and morphological annotations was found to be non-inferior to embryologists but had no significant impact on continuing pregnancy or implantation rates (Ahlström *et al.*, 2022).

There is little doubt in the IVF community that the human eye is a limiting factor when assessing the microscopic phenotypic manifestations of embryogenesis. In recent years, there have been numerous attempts to standardize embryo selection with artificial intelligence (AI), including subfields such as machine learning and deep learning (DL), using visual data to develop more objective embryo assessment techniques. Such endeavors have focused on different technologies and outcome predictions, such as blastocyst formation using morphokinetic annotations (Petersen *et al.*, 2016), embryologist decisions to transfer and cryopreserve embryos (Bormann *et al.*, 2020) and implantation outcome (Tran *et al.*, 2019; VerMilyea *et al.*, 2020; Wang *et al.*, 2021). Models based on visual inputs that are designed to predict implantation outcome possess the greatest potential as they do not rely on traditional, subjective inputs, namely manual morphological assessment or morphokinetic annotations. The models are instead trained using still images or TLI videos of embryos with known implantation data (KID), which are then used to objectively identify patterns associated with success. The detected patterns may include previously identified morphological and morphokinetic features, but the high-throughput nature of AI methods also affords them the ability to identify factors that manual embryo study cannot.

A common metric to assess the performance of such tools is the AUC of a receiver operating characteristic (ROC) curve (Kragh and Karstoft, 2021). This allows interpretation of a model's ability to distinguish between outcomes in a binary classification system and is calculated by plotting true positive rate (sensitivity) over false positive rate (1-specificity). Results vary depending on the primary outcome of the studies. Models attempting to predict embryo quality, defined by standard morphological assessment, perform extremely well with AUCs as high as 0.98 (Khosravi *et al.*, 2019). Several papers have been published in recent years assessing the ability of AI models to predict implantation, with AUCs promisingly sitting around the 0.7 mark (Kragh *et al.*, 2019; Chavez-Badiola *et al.*, 2020b; Ueno *et al.*, 2021; Berntsen *et al.*, 2022; Loewke *et al.*, 2022). Given that euploid embryos implant on average ~70% of the time (Tiegs *et al.*, 2020), it is likely that these results are approaching maximum performance when considering embryo factors alone.

Although results thus far have been promising, there is yet again a lack of data validating their clinical performance (Storr *et al.*, 2018). One important hurdle is software generalization: the ability of AI to perform consistently regardless of known differences such as embryo culture methods, imaging hardware and patient populations. Such difficulties in generalization are largely due to the tendency of AI models to overfit their training data (Berntsen *et al.*, 2022). Generalization is theoretically achievable with DL through the use of large, heterogeneous datasets and advancements in DL techniques; however, this is logistically difficult (Zhang *et al.*, 2016; Neyshabur *et al.*, 2017). Therefore, human morphological assessment remains the gold standard for embryo selection as the AI community continues to debate the challenges and merits of a generalized versus clinic-specific approach.

The objective of this study was to assess the implantation prediction agreement between a large cohort of embryologists grading TLI videos of embryos developed to the blastocyst stage with KID, and to compare their implantation prediction performance against a purpose-built DL model.

Materials and methods

Study design

In this prospective study, 39 embryologists were recruited from 13 different countries in April and May of 2020. The aim was to assess the inter- and intraobserver agreement of embryologists when asked to predict the likelihood of implantation for blastocysts based on retrospectively collected TLI videos of developing embryos and then compare their performance with that of a purpose-built deep neural network (DNN) model.

The dataset initially consisted of 143 retrospectively collected TLI videos of transferred embryos cultured in EmbryoScope or EmbryoScope+ TLI systems (Vitrolife, Aarhus, Denmark) at two different IVF clinics from cycles performed between 2018 and 2019. Videos were 500 by 500 pixels, with time-lapse intervals of 15 min. Mean patient age was 35.7 years (age range 26–44 years). Both ICSI and IVF were utilized in the fertilization process (98% and 2%, respectively), using 141 autologous oocytes and 2 donor oocytes. Inclusion criteria for this experiment were: embryos with KID, determined using ultrasound evidence of a gestational sac at 5–8 weeks as identification of implantation (Zegers-Hochschild *et al.*, 2017); TLI culture and image capture until at least 90-h post-insemination (HPI); and blastocyst stage embryo, determined by senior embryologists. Overall, seven videos were excluded from model training and final analyses either because the embryos had not reached the blastocyst stage (n = 6) or because the implantation outcome was unclear (n = 1), resulting in a final dataset of 136 videos (Fig. 1). Incubation times of the final 136 embryos ranged from approximately 95 to 144 HPI, with one embryo cultured to Day 4, 123 embryos cultured to Day 5 and 12 embryos cultured to Day 6. The final ratio of negative to positive implantation was approximately 1:1 (52.3% no implantation (n = 71) versus 47.7% implantation (n = 65)). Transfers were either single embryo transfers with known implantation outcomes or double embryo transfers where implantation outcome was guaranteed through virtue of deduction (i.e. implantation rate either 100% or 0%). PGT was not performed on any embryo. There was one recorded incidence of monozygotic twins. OP and birth outcomes were only available for a subset of the data and were therefore not included in the analysis.

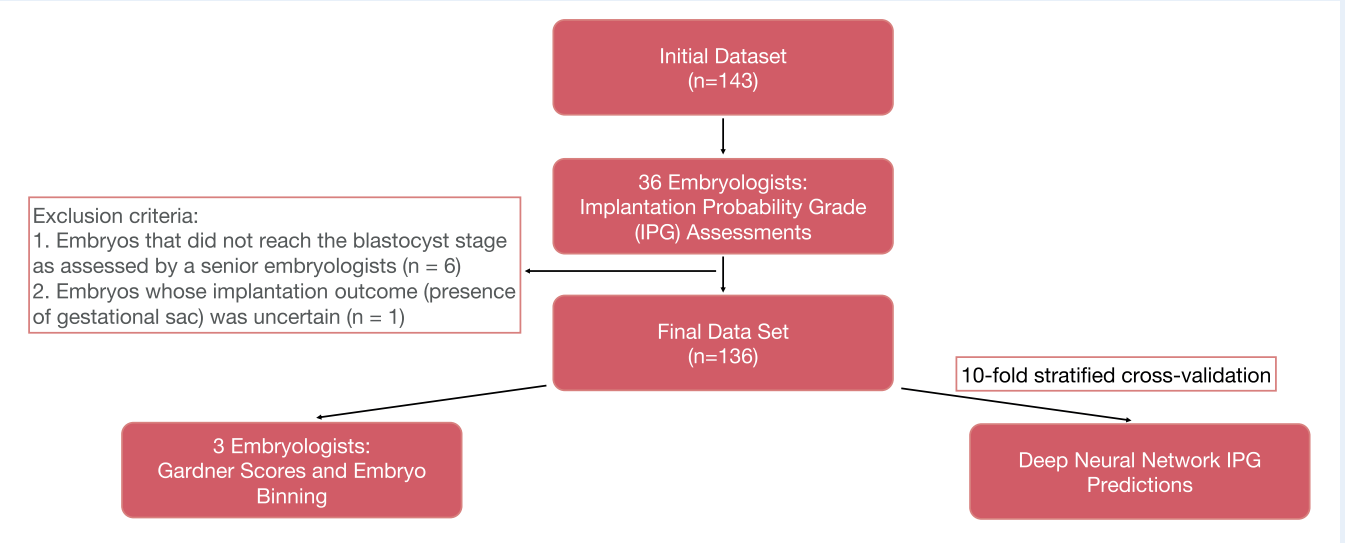


Figure 1. Flow chart detailing the study design. Summary workflow detailing the main steps of the study. After results were collected, statistical analysis was performed in order to assess embryologist agreement and to compare embryologist and DNN model performance.

Embryologist implantation prediction grading

We asked 37 embryologists to provide TLI implantation predictions for this study based on their analysis of entire TLI videos. They were not required to provide morphological scores or morphokinetic annotations. One participant failed to complete the study, leaving a total of 36 embryologists for analysis. Three video sets, each containing a randomized subset of 100 TLI videos from the original 143, were generated and uploaded to a shared folder. Participants were given temporary permission to view the set. They were required to complete each set before continuing to the next set, which became available after seven days. Participants were blinded to the fact that videos repeated between sets. Participants were asked to provide an implantation probability grade (IPG) between 1 and 5 for each embryo with descriptions as follows: 1—Very low likelihood of implantation; 2—Low likelihood of implantation; 3—Moderate likelihood of implantation; 4—High likelihood of implantation; 5—Very high likelihood of implantation. These results were collected in an Excel proforma provided to each participant with the release of each set. No additional clinical information was provided. At the end of the experiment, participants completed a questionnaire containing questions relating to years of experience, job title, embryo grading system(s) used at their clinic, previous experience with TLI hardware and, if so, which systems.

Analysis of each embryologist's IPG predictions was conducted on 42 embryos graded once, 39 graded twice and 55 graded three times. A total of 10 260 individual predictions were made (285 per embryologist). Sample sizes were deemed adequate after reviewing relevant literature comparing inter- and intraobserver variability (Storr *et al.*, 2017; Hammond *et al.*, 2020).

Blastocyst morphological grading

Two additional embryologists were recruited from independent clinics to provide Gardner scores for all embryos that met the inclusion criteria. In the event that their grades did not agree, a senior embryologist from the initial prediction study acted as a tie breaker. Each of the three embryologists providing Gardner scores had at least 5 years' experience. After Gardner scores were finalized, embryos were divided into three groups: top quality embryos graded 3-6AA (n = 32), fair quality embryos graded 3-6AB, BA or BB (n = 59) and poor quality embryos graded 1–2 for expansion or a C for either inner cell mass and/or trophectoderm (n = 45) (Fig. 2) (Gardner *et al.*, 2000; Cetinkaya *et al.*, 2015; Petersen *et al.*, 2016). The percent of embryos that achieved implantation in each group remained relatively balanced (59.4%, 42.4% and 46.7%, respectively).

Model architecture

Two separate DNN models were used to generate IPGs. For the first model, five frames were extracted from each video as the model input based on the expected optimal timings of developmental stages described in the Istanbul consensus (Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology, 2011). Resulting HPIs were 28, 44, 68, 92 and 116, corresponding to the ideal timings of the 2-cell, 4-cell, 8-cell, morula and blastocyst stages, respectively. These frames were fed through a pretrained EfficientNetV2 encoder (Tan and Le, 2021) to extract content features. Extracted frames were linearly resized to 500 × 500 pixels and normalized to standard values for the EfficientNetV2 pretrained network which concatenates the resulting encoded vectors and fine-tunes a linear layer for classification. For the EfficientNet model, the initial learning rate was 1e−5, with a dropout rate of 0.1. Extracted data were used as input for a classification network which outputs IPGs.

For the second model, 100 frames were extracted from each video as the model input (from 20 to 120 HPI at 1-h intervals) (Fig. 3). These sample frames were encoded using the 'Embryo morphology encoder' (Fig. 4), a conventional neural network encoder based on the

Seg-Net architecture (Badrinarayanan *et al.*, 2017). It was trained on semantic segmentation of embryo components from time-lapse frames (images) and was designed to take TLI frames and output vectors encoding data related to visual elements of the developing embryo.

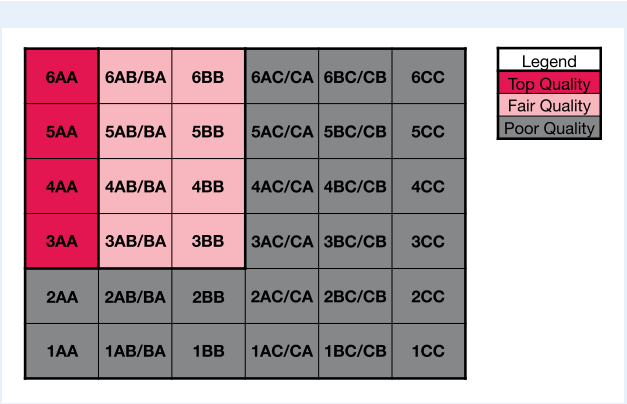


Figure 2. Binning of embryos into quality groups using the Gardner score. Each TLI video was graded using the Gardner criteria by a panel of embryologists. Embryos were then binned into top, fair and poor quality embryos based on their Gardner score (Gardner *et al.*, 2000; Cetinkaya *et al.*, 2015; Petersen *et al.*, 2016). TLI, time-lapse imaging.

A positional embedding was then concentrated to each of the encoded vectors before being fed through a transformer encoder. For the Transformer-based model, a learning rate of 5e−5, and a dropout rate of 0.3 were used.

In parallel to the transformer, one frame from HPI 110–120 was fed through the pretrained EfficientNetV2 encoder with the aim of extracting content features specific to the embryo's blastocyst stage of development. The transformer block and the pretrained EfficientNetV2 were followed by a multilayer perceptron head which output the model predictions for which a classification loss was minimized during the training. The optimizer used was Adam (Kingma and Ba, 2017).

Both models were combined into a DNN ensemble to output a single IPG from both models' predictions.

Ten-fold cross-validation

The TLI videos were randomly divided into 10 equal folds with implantation: no implantation ratios of 1:1. In total, 10 iterations were carried out covering the entire 136 video dataset. Within each iteration, 7 folds were selected as the training set, with 2 folds for validation and 1 fold for testing. At the conclusion of the series, the stopping epoch for each run for the test fold was chosen based on the highest AUC value of the validation folds.

Several intrinsic factors were anticipated to artificially affect the classification label, including brightness, light source and shadow of the

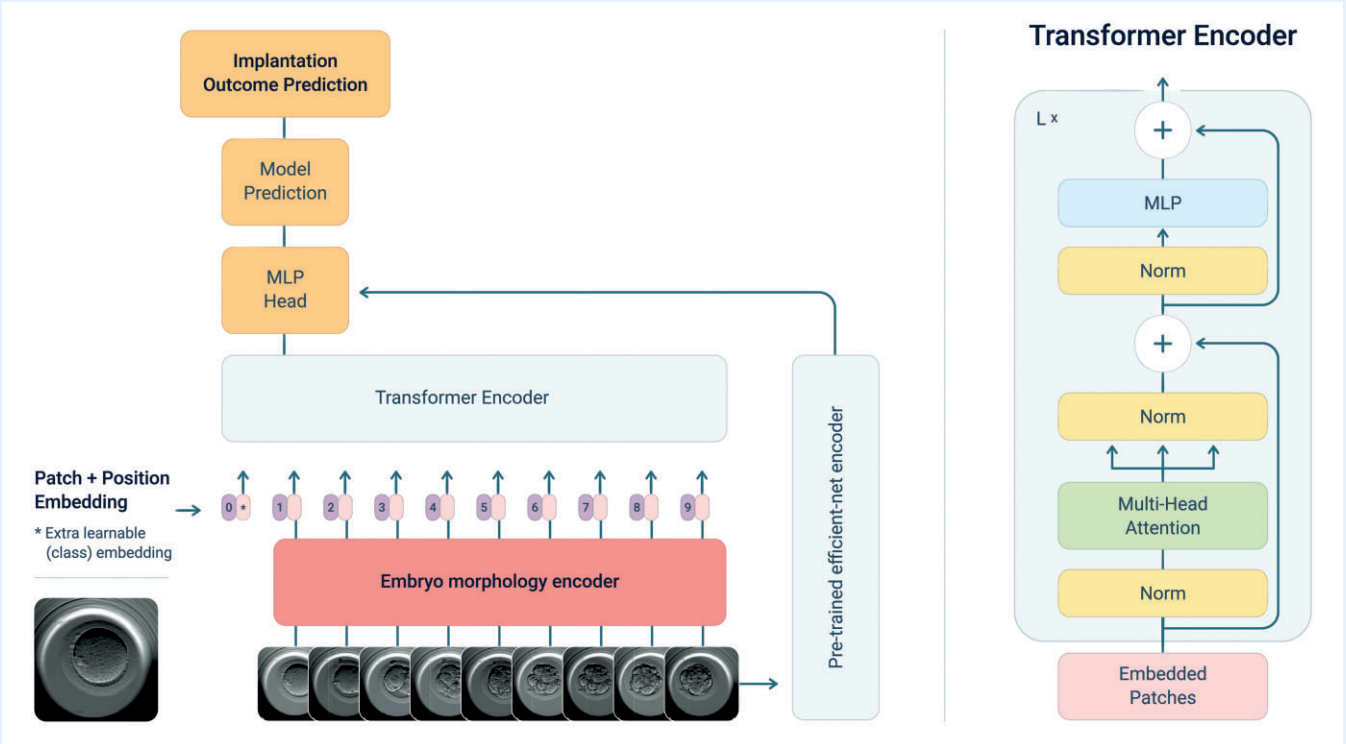


Figure 3. Transformer-based deep neural network model overview. Summary of model 2. A video is sampled uniformly into a fixed length of 100 frames, each frame is encoded, positions are embedded, and the resulting sequence of vectors is fed to a standard transformer encoder. One of the last frames was fed to a pre-trained efficient-net encoder (EfficientNetV2), the output vectors from EfficientNetV2 and the transformer were concatenated and passed through a multilayer perceptron (MLP) which output the model predictions. Image adapted from Figure 1 of Dosovitskiy *et al.* (2020).

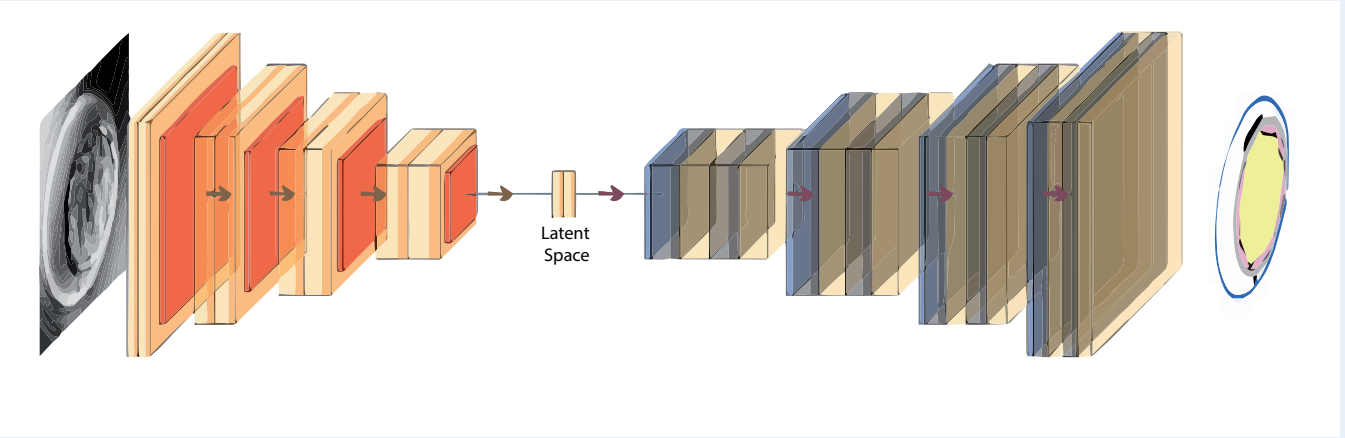


Figure 4. Embryo morphology encoder overview. An auto-encoder architecture based on Seg-Net (Badrinarayanan *et al.*, 2017) where the skipped connections were removed to encode embryo morphology by training the model on a semantic segmentation task.

EmbryoScope model, embryo position within the well and zoom scale (Supplementary File S1). In order to enhance the model's robustness to such factors and prevent overfitting, augmentations were applied to the training set including rotations, scaling and global color adjustments (Supplementary File S1).

Raw predictions from each fold were between 0 and 1. The optimal prediction threshold was determined based on accuracy over the validation folds, binning thresholds for the transformation to IPGs of 1 to 5 were calculated by normalizing raw predictions to the extended range.

Statistical analysis

To examine the impact of the embryologists' characteristics on implantation prediction accuracy, binomial regression was performed (71 cases of non-implantation and 65 cases of implantation). For the characteristic 'country', three cohorts were generated loosely based on geographical territories: embryologists working in the USA, Western Europe and Turkey, and the rest of the world. CIs of 95% were applied to odds ratios.

The Fleiss kappa (κ) coefficient (Fleiss, 1971) was used to assess the level of agreement between the 36 embryologists. In total, three kappa scores were calculated according to the following groupings: fair quality embryos (n = 59), poor and top quality embryos combined ('poor+top quality group', n = 77) and all embryos. Like Hammond *et al.* 2020, a combined poor+top group was selected for kappa analysis as a presumption was made that embryologists would have stronger agreement for these embryos. The suggested kappa value interpretation is: 0.0–0.20 poor agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 good agreement and 0.81–1.00 very good agreement (Altman, 1990). To evaluate statistical significance between Fleiss kappa coefficients, Z-test approximation was used.

The three test sets provided to the embryologists contained a number of TLI videos that were repeated two or three times across sets. A test–retest agreement method using intraclass correlation (ICC) was calculated to determine how consistent the embryologists were with themselves, based on the embryos that were graded either two or three times. ICC values may be interpreted as follows: >0.8 = high-perfect agreement; 0.7–0.8 = good agreement; 0.5–0.6 = regular-moderate agreement; 0.2–0.4 = fair agreement; and 0.0–0.2 = poor agreement (Shrout and Fleiss, 1979).

To allow for performance assessment, which requires binary inputs, the five-point IPG scale was divided into two groups. IPGs of 3, 4 and 5 were considered predictions of implantation, whereas IPGs of 1 and 2 were considered predictions of no implantation. The IPG threshold of 2 was selected because it resulted in the highest overall interobserver agreement for the embryologists (data not shown). The embryologists' and model's performances were then evaluated by the following statistical measures: accuracy, F1, sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV). For these comparisons, the bootstrapped method for the mean was calculated for each metric by resampling the 36 embryologists' performances 10 000 times, and the 95% CI to the mean was created for each measure.

The Kruskal–Wallis test was applied to evaluate whether a statistical difference was present between the three quality groups. The Mann–Whitney *U* test was then used to assess significance between individual embryo quality groups. Comparisons were also made between the model and embryologists using ROC curves, with AUC used for performance comparison. AUCs were calculated for the overall dataset, and after binning into Gardner quality groups.

Results

Characteristics of embryologist cohort and their effect on accuracy

The characteristics of the embryologists who participated in the study are summarized in Table I. In total, 36 embryologists working in 12 different countries and 29 different IVF clinics provided implantation predictions for this study. At the time of the study, 27.8% of participants practiced in North America (10/36), with 55.6% in Europe (20/36),

Table I Characteristics of participating embryologists.

Embryologists numbers	Country	Years active	Embryo grading	Specific TLI experience	Qualifications	Job title (self-reported)
1	USA	32.00	a, f	0	PhD	Consultant Embryologist
2	USA	20.00	a	1,2	Masters	Senior Embryologist/Lab Supervisor
3	USA	16.00	a	0	Masters	Senior Embryologist/Lab Supervisor
4	USA	9.00	a	1,3,4	Masters	Senior Embryologist
5	USA	4.00	a	0	PhD	Embryologist
6	USA	3.00	a	1	Masters	Embryologist
7	USA	3.00	b	1	Masters	Embryologist
8	USA	3.00	a, b	0	Phd	Senior Embryologist
9	USA	3.00	a	0	Masters	Embryologist
10	USA	1.00	f	0	Masters	Junior Embryologist
11	Spain	15.00	a	1,3	Masters	Senior Embryologist
12	Spain	6.00	a, c	1,3	Masters	Senior Embryologist
13	Spain	4.00	c	1	Masters	Embryologist
14	Spain	2.00	a	1,2	PhD	Embryologist
15	Spain	2.00	c	1,2,3	2 Masters	Embryologist
16	Spain	0.25	c	1,2	2 Masters	Embryologist
17	Ukraine	9.00	f	1,5	Masters	Senior Embryologist
18	Ukraine	4.00	a	0	PhD	Embryologist
19	Ukraine	3.00	a	0	BSc	Embryologist
20	Ukraine	1.00	a	0	Masters	Junior Embryologist
21	Ukraine	0.50	c	Unknown	Masters	Junior Embryologist
22	India	8.00	a	1	Masters	Senior Embryologist/Lab manager
23	India	8.00	d	1,5	PhD	Consultant Embryologist
24	India	6.00	a, e	1	Masters	Senior Embryologist
25	India	2.00	a, d	0	Masters	Junior Embryologist
26	UK	5.00	a	1,2	2 Masters	Embryologist
27	UK	2.00	a	1,2	Masters	Embryologist
28	Turkey	14.00	f	1	Masters	Senior Embryologist
29	Turkey	9.00	d	1	Masters	Senior Embryologist/Lab Director
30	Bulgaria	13.00	a	1	Masters	Senior Embryologist
31	Bulgaria	8.00	a	0	PhD	Embryologist
32	UAE	8.00	f	6	Masters	Senior Embryologist
33	Lithuania	19.00	a, f	6	PhD	Senior Embryologist
34	Ghana	2.00	a	0	BSc	Embryologist
35	Georgia	10.00	a	0	Masters	Consultant Embryologist
36	Czech	6.00	a	1,3,5	DR Rer Nat	Clinical Embryologist

TLI, time-lapse imaging.
Embryo Grading: [a]Gardner, [b]SART, [c]ASEBIR, [d]Istanbul Consensus, [e]KIDScore D5, [f]Local grading system.
TLI Experience: [0]None, [1]Embryoscope/Embryoscope+, [2]Geri, [3]PrimoVision, [4]Eeva, [5]MIRI TL.

13.8% in Asia (5/36) and 2.8% in Africa (1/36). Years of experience (post-undergraduate qualification) ranged from 3 months to 32 years, with a mean of 7.1 years (standard deviation 6.8 years). Female to male ratio was exactly 2:1. The majority of the participants reported previous or current experience with TLI incubators (69%, 25/36), with EmbryoScope/EmbryoScope+ being the most commonly used device. Unsurprisingly, the embryo grading system used by participants had a regional distribution. The Gardner score was the most ubiquitous system (67%, 24/36), showing universal distribution. ASEBIR was particularly prevalent amongst embryologists working in Spain (4/6), and 17% (6/36) of participants routinely used two different grading systems, for which the Gardner score always featured. Most (94%) participants had at least a master's degree (34/36) with 25% (9/36) having completed a doctorate.

Logistic regression comparing odds ratios showed no statistically significant differences (at $\alpha = 0.05$) between implantation prediction

Table II Logistic regression comparing odds ratios between implantation prediction accuracy and embryologist characteristics.

	Successes/total		
Predictors	**Odds ratios**	**95% CI**	***P***
(Intercept)	1.10	0.92–1.31	0.311
Country [United States of America]	1.00	0.86–1.16	0.970
Country [Western Europe + Turkey]	0.98	0.84–1.14	0.798
Degree [PhD]	1.03	0.90–1.18	0.639
Embryo Scoring System [with Gardner score]	1.04	0.91–1.20	0.537
Experience in TLI [Yes]	1.04	0.89–1.21	0.611
Log Years of Experience	0.96	0.89–1.04	0.286

No characteristic predicting variable was statistically significant at $\alpha = 0.05$.
TLI, time-lapse imaging.
The baseline ('intercept') was an embryologist with the predictors: Country [Other], Embryo Scoring System [without Gardner], Experience in TLI [No] and Degree [No PhD].

accuracies and embryologist characteristics (Table II). To assess the effect of characteristic variables on embryologist accuracy, a binomial regression model was built where the response variable was the number of correct predictions (true positives and true negatives) each embryologist provided in the 136-video set. The variables Country, Degree, Embryo Scoring System and Experience in TLI were indicator variables in this regression. The baseline ('intercept') was an embryologist with the predictors: Country [Other], Embryo Scoring System [without Gardner], Experience in TLI [No] and Degree [No PhD].

Interobserver and intraobserver agreement between embryologists

Figure 5 shows the interobserver agreement, which was assessed using weighted kappa scores. Embryos classified as poor and top quality after Gardner grading were grouped together (n = 77) and compared to embryos classified as fair quality (n = 59), and to the entire set of 136 embryos (Fig. 5). The kappa score for the combined poor+top quality group (weighted κ: 0.65; 95% CI: 0.6–0.7) indicated good agreement between embryologists. In contrast, the interobserver agreement for the fair quality group only showed fair agreement (κ: 0.25; 95% CI: 0.18–0.321), significantly lower than that of the poor+top quality group ($P < 0.001$). Overall agreement was moderate (weighted κ: 0.56; 95% CI: 0.51–0.61).

ICCs measuring the test–retest reliability, a measure of intraobserver agreement, were calculated for TLI videos graded twice (n = 39; ICC: 0.77; 95% CI: 0.64–0.91) and three times (n = 55; ICC: 0.75; 95% CI: 0.67–0.89) by each embryologist, showing good consistency (Supplementary Table SI).

Embryologist characteristics were found to not statistically affect the agreement. For embryologists working in the USA the kappa was 0.51 (95% CI: 0.45–0.57, n = 10), for Western Europe and Turkey $\kappa = 0.52$ (95% CI: 0.47–0.58, N = 14), and for Other $\kappa = 0.51$ (95% CI: 0.45–0.56, n = 12). For embryologists with a PhD, $\kappa = 0.49$ (95% CI: 0.43–0.55, N = 9) compared to $\kappa = 0.55$ (95% CI: 0.50–0.60, n = 27) for those without. For embryologists with up to 5 years of experience $\kappa = 0.55$ (95% CI: 0.50–0.60, n = 19), for embryologists with 5–10 years of experience $\kappa = 0.51$ (95% CI: 0.45–0.56, n = 12), for embryologists with 10–15 years of experience $\kappa = 0.41$ (95% CI: 0.32–0.50, n = 3), for embryologists with over 15 years of experience $\kappa = 0.48$ (95% CI: 0.41–0.55, n = 4).

Performance of embryologists versus DNN model after binary binning according to implantation prediction

To calculate implantation prediction accuracy, an IPG of ≤ 2 was considered a prediction of failure to implant, whereas ≥ 3 was considered a prediction of successful implantation. Figure 6 shows the distribution of the IPGs for each embryo (n = 4896). For TLI videos that were graded multiple times across the video sets, only the first recorded IPG was analyzed. Overall, participants appeared to over-estimate the likelihood of implantation with 75.6% of embryologists' IPGs predicting implantation (n = 3699) versus 24.4% predicting no implantation (n = 1197), despite the balanced dataset.

In order to assess the performance of the DNN, statistical analysis was conducted on IPGs generated for the dataset and compared to those of the embryologists. The DNN performed better than the embryologists in every key metric measured, with a 20.5% increase in accuracy (62.5% vs 51.9%), 10.3% increase in F1 (67.1% vs 60.8%), 1.7% increase in sensitivity (80% vs 78.7%), 21.8% increase in NPV (71.7% vs 58.9), and 16.1% increase in PPV (57.8% vs 49.8%) (Fig. 7). Excluding sensitivity, the improvement in performance achieved for the benchmark metrics by the DNN model was found to be significant ($\alpha = 0.05$). All individual values with 95% CIs and SDs for the DNN model are shown in Supplementary Table SII.

Performance of embryologists versus DNN model after binning according to Gardner groups

Figure 8 compares the implantation prediction accuracies for embryologists with those of the DNN model after embryos were binned into Gardner groups. Using the Kruskal–Wallis test and then evaluating significance with the Mann–Whitney *U* test, a significant difference was found between embryologist accuracy when predicting implantation for fair embryos (44.6%, 95% CI 36.2–52.9) compared to poor embryos (57.4%, 95% CI 49.1–65.7) ($P < 0.0001$) and top embryos (57.5%, 95% CI 49.2–65.8) ($P < 0.0001$). There was no statistically significant difference found between poor and top embryo prediction accuracies. The DNN model achieved higher accuracies for the poor (62.2%), fair (61.0%) and top (65.6%) embryo quality groups compared with the embryologists.

ROC curves showing true positive rate versus false positive rate are shown in Fig. 9. The AUC for the DNN was higher overall (0.70 DNN vs 0.61 embryologists) as well as in all the Gardner groups (DNN vs embryologists for Poor: 0.69 vs 0.62; Fair: 0.67 vs 0.53; Top: 0.77 vs 0.54). Multiclass analysis showed that the DNN model performed better than every embryologist when ROCs were plotted individually (Supplementary Fig. S1).

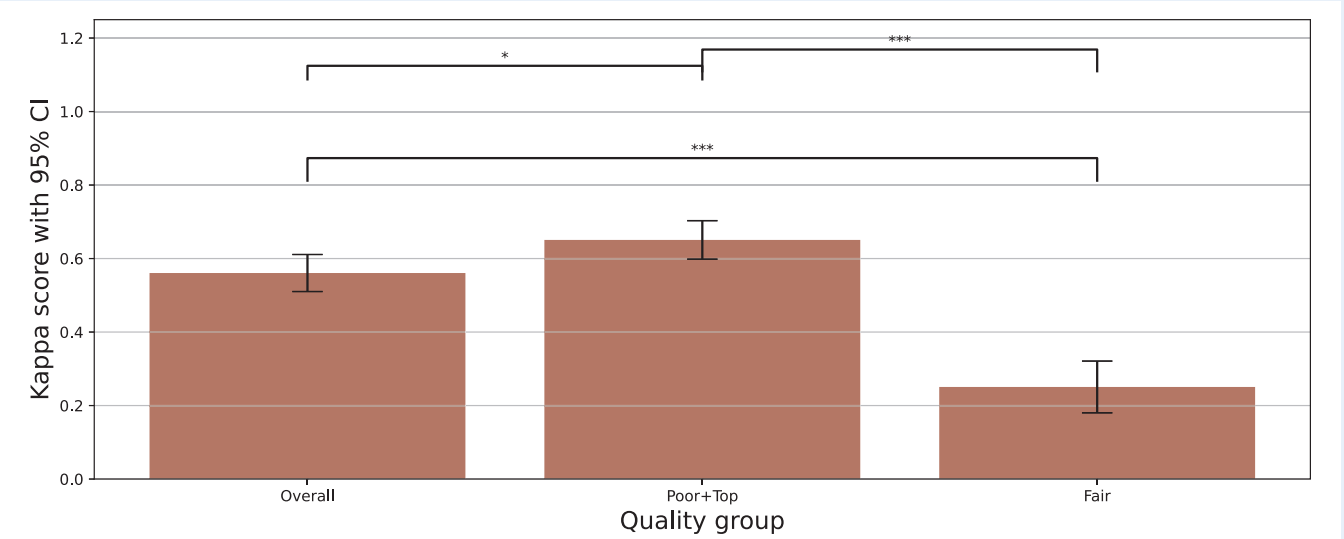


Figure 5. Weighted kappa scores comparing embryologist agreement according to Gardner grouping. Embryologists were significantly more likely to agree on implantation probability grades (IPGs) for embryos in the poor + top group when compared with fair embryos or the overall data. ***P-value < 0.001; *P-value < 0.05.

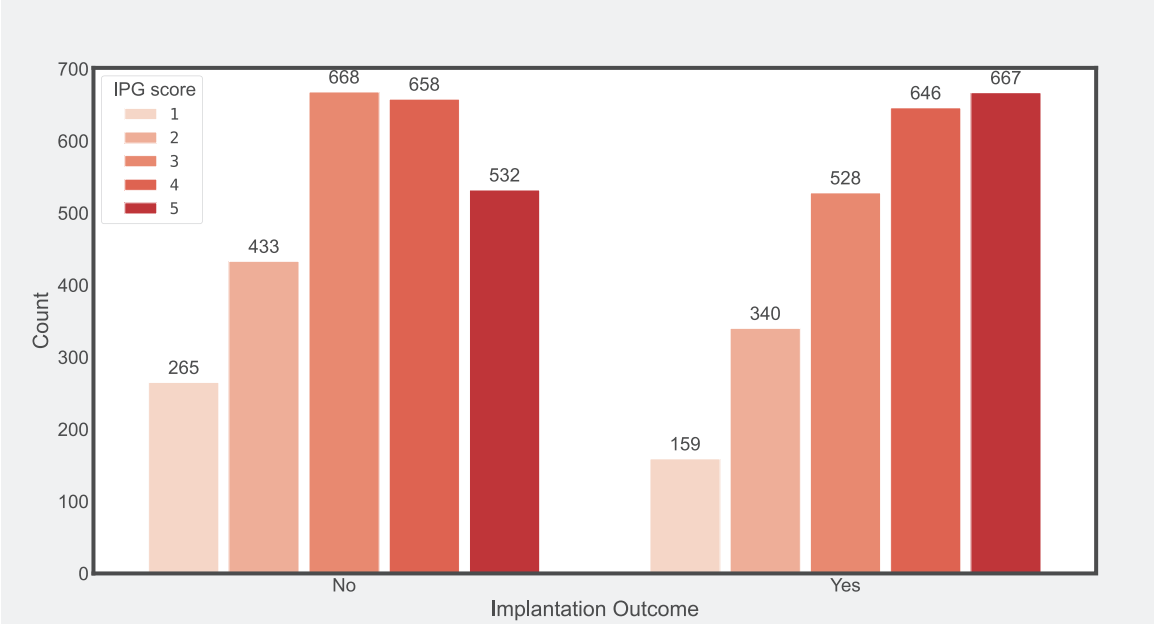


Figure 6. Distribution of embryologist implantation probability grade (IPG) predictions per actual implantation outcome.

Discussion

Studies assessing inter- and intraobserver agreement in IVF have traditionally been performed using data and embryologists from a single center, using static images of embryos and tend to either compare specific embryo quality scores or embryo selection (Arce *et al.*, 2006; Storr *et al.*, 2017; Hammond *et al.*, 2020). In this study, we assess the underlying factors affecting both embryologist agreement and outcome prediction in a novel way. To the best of our knowledge, this is the largest and most diverse cohort of embryologists assembled for a study of this kind. The use of TLI videos and the prediction of

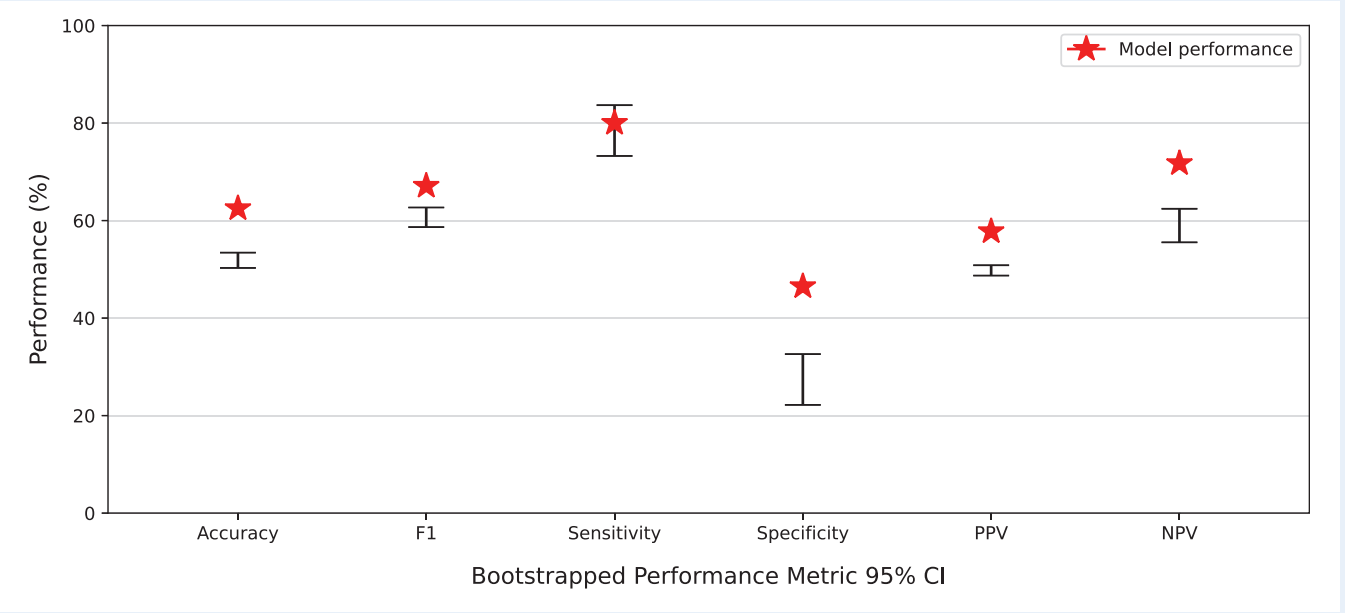


Figure 7. Performance metrics of embryologists compared to the deep neural network (DNN) model. The DNN model (red star) outperformed the study population in a number of key metrics. See Supplementary Table SII for figures.

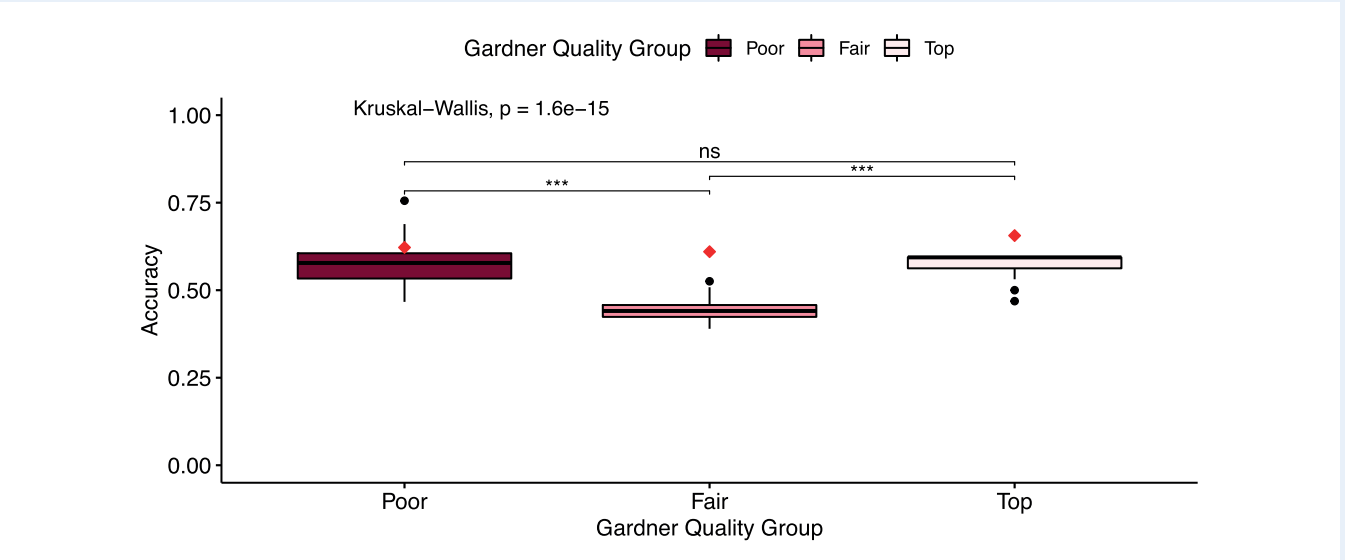


Figure 8. Deep neural network (DNN) model and embryologist implantation prediction accuracies after Gardner quality grouping. Red diamonds represent DNN model accuracies. Black circles represent individual embryologist outliers. After performing the Kruskal–Wallis test, embryologists were significantly more accurate at predicting implantation for embryos deemed poor and top quality according to Gardner quality grouping. There was no statistically significant difference found between accuracies for the poor and top group. To determine whether the DNN model was more accurate, the approximate one sample Z-test for proportion was used. It was found that for the fair group, the model was more accurate ($P = 0.0245$). ***P-value < 0.001. ns, non-significant.

implantation are also relatively unique in the assessment of observer agreement. Ultimately, the results demonstrate embryologist adaptability whilst allowing direct comparison of their performance with that of a DNN model trained for the purpose of implantation prediction.

Initial assessment of embryologist predictions showed that none of the characteristics recorded in this study significantly affected their accuracy or level of agreement. These include country of residence, years of experience, embryo score generally used in clinical practice,

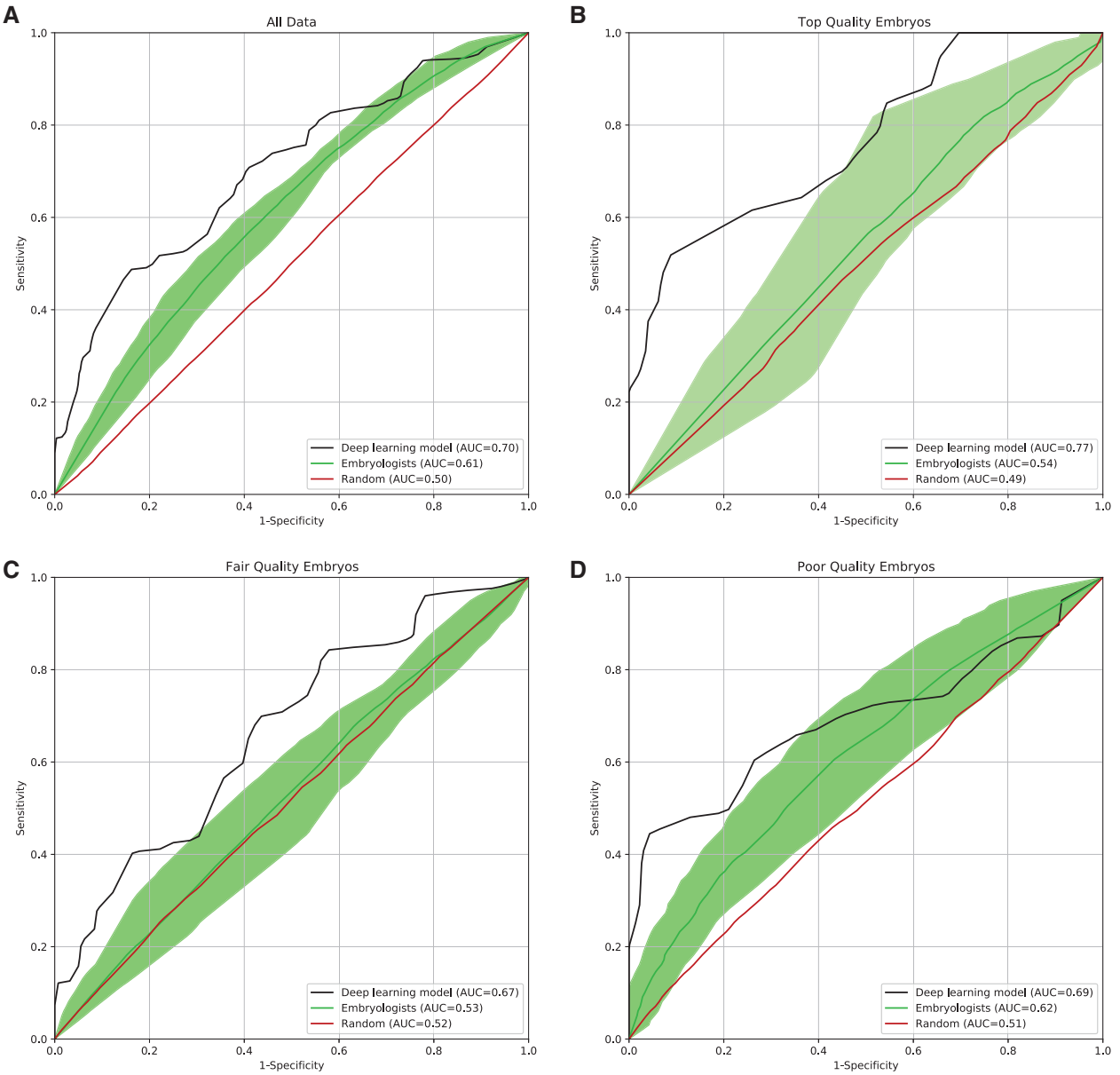


Figure 9. ROC curves with AUCs comparing DNN and average embryologist implantation prediction performance for the full dataset and each of the Gardner quality groups. The DNN outperformed the study population overall and in each Gardner group. ROC, receiver operating characteristic (sensitivity over 1-specifivity). AUC, area under the curve, the cumulative distribution function.

previous TLI experience and educational degree. This lack of significant association between interobserver agreement and years of experience (Bendus *et al.*, 2006; Storr *et al.*, 2017), credentials (Bendus *et al.*, 2006) and grading system (Hammond *et al.*, 2020) has been previously observed in multiple studies. To our knowledge, there is a lack of published data from multicenter studies assessing the correlation between agreement and country of residence. As previously discussed, due to the subjective nature of embryo assessment, comparative studies between embryologists have tended to focus on agreement. Thus, the findings of this study that characteristics did not have an association with accuracy should be further investigated in future studies.

Interobserver agreement varied depending on the quality of the embryos being assessed. Embryologists were more likely to agree on implantation probability for top and poor quality embryos ($\kappa = 0.65$) than for fair quality embryos ($\kappa = 0.25$). This is somewhat explainable as, with the use of TLI, embryos exhibiting morphokinetic events associated with poor prognosis (e.g. direct or indirect cleavage) and good prognosis (e.g. appropriate event timing) are likely to stand out to the trained embryologist as poor and top quality respectively, whereas fair

quality is a more ambiguous description (Rubio et al., 2012; Liu et al., 2014). This data supports the findings of Hammond et al. (2020) in which a vastly inferior kappa score was observed for borderline blastocysts compared to top quality and non-viable blastocysts when embryologists were tasked with selecting embryos for cryopreservation. Hammond hypothesizes that the subjective nature of current manual grading systems inherently leads to reporter inconsistency, particularly for borderline blastocysts, thus negatively impacting patient care (Hammond et al., 2020). Farin et al. (1995) also observed inferior agreement between embryos given intermediate grades compared to top and bottom grades when assessing bovine embryo quality (Farin et al., 1995). Storr et al. (2017) observed that when embryologists were asked to select a single day 5 embryo for transfer (i.e. highest quality embryo in a cohort), the overall kappa scores were comparable to that of the poor + top quality group in our study. Intraobserver agreement was also similar, with overall good agreement (Storr et al., 2017). These findings, despite exploring somewhat different outcomes, support our conclusion that traditional morphological grading leads to inconsistencies in embryo selection, for fair/borderline embryos.

It is this subjectivity, well recognized in the ART community (Paternot et al., 2009; Sundvall et al., 2013; Rocha et al., 2016), which warrants the development of techniques that utilize a more objective outcome, such as implantation, when creating assistive AI solutions for ART objectives (Khosravi et al., 2019; Tran et al., 2019; Ueno et al., 2021). While models designed to mimic the classical approach of grading embryo quality, invariably using predefined morphological metrics, and essentially reproducing manual assessments, could be useful for reducing embryologist workload, a model that can predict the chances of a positive pregnancy could potentially lead to improved clinical outcomes, thereby conferring additional benefits to both the IVF clinics and their patients.

When compared with the DNN model, both the embryologists and the model were equally sensitive; however, the embryologists had significantly lower specificity, PPV, NPV and overall accuracy. Importantly, the DNN model performed similarly in all quality groups, not showing the dip in accuracy for the fair group that was observed amongst embryologists. It is likely that this reflects the more objective approach of the model, which was trained on a relatively balanced dataset with regards to embryo quality. As this is the cohort that morphological grading appears to find challenging to predict, it represents a justified area of focus for AI assistance in order to aid embryologists in their clinical decision-making (Rocha et al., 2016).

Comparing the AUCs of the ROC between the DNN model and embryologists, not only did the model outperform the average embryologist (0.70 and 0.61, respectively), it performed better than each of the 36 embryologists individually (Supplementary Fig. S1). Importantly, the model showed superior performance for all Gardner quality groups. Kragh et al. (2019) reported no significant difference between the AUCs of an implantation prediction model and embryologists (0.66 and 0.64, respectively). However, the implantation rate of their dataset was 29%, and much of the criticism of AUC as a metric for performance in this field lies in the fact that it is susceptible to bias in the event of uneven datasets. Tran et al. (2019) reported a very impressive AUC of 0.93 for their DL implantation prediction model. However, these results seem unrealistic and have been disputed due to the labeling of non-transferred embryos with negative implantation, which made up the vast majority of their dataset and likely had a large impact on results (Chavez-Badiola et al., 2020a; VerMilyea et al., 2020). VerMilyea et al. used convolutional neural networks (ResNet-152 and DenseNet-161) designed to predict implantation for static images of day 5 blastocysts. The AUC was reported as 0.71; however, they acknowledge imbalance of their test set as a limiting factor (approximately 1:3 implanted to non-implanted) (VerMilyea et al., 2020).

More recently, Loewke et al. (2022) published an impressive study concerning a static image-based AI model also designed to predict implantation probability. They reported AUCs of 0.74 for all embryos, 0.64 for non-PGT transferred embryos, and 0.63 for transferred euploid embryos. Non-transferred aneuploid embryos, confirmed using PGT, were included in the training of the algorithm. This to some extent addresses the problem of generalizing AI models for non-transferred embryos; however, the issue of viable mosaic embryos being excluded from training was not discussed. Additionally, to generate AUCs comparing model performance with that of the embryologists, they converted standard of care alphanumeric morphokinetic grades to a 36-point numerical system. While this approach allows for comparison that does not require embryologists to grade embryos in a manner they are not familiar with, it makes it harder to objectively compare AUCs, as they are assessing different clinical endpoints (i.e. embryologists' morphological assessments compared to the model's implantation predictions).

In our study, 75.6% of embryos were predicted by embryologists to result in implantation. This is in spite of the fact that implantation rates are seldom this high. For example, the overall implantation rate for transferred embryos was ~39.8% in the USA in 2019 according to the CDC national summary (2019 Assisted Reproductive Technology Fertility Clinic and National Summary Report | CDC, 2022). In our study, the implantation rate for the dataset was 47.7%. The difference in predicted implantation compared to actual implantation rate suggests that embryologists tend to overestimate the likelihood of implantation. This overestimation of clinical outcome (this optimism) is a phenomenon seen across many aspects of clinical decision-making, and indeed decision-making in general, which could be considered analogous to the 'framing effect' described in decision-making theory (Tversky and Kahneman, 1981). This study essentially poses the question 'is this embryo good enough to implant?', defining the terms of success in a slightly different clinical framework than embryologists are used to. Putting the conscious observer into an experimental situation, suddenly the question has been framed relative to implantation. A correct prediction of implantation (true positive) is the most desirable outcome, followed by correct non-implantation (true negative). The least desirable outcome according to frame theory is false non-implantation (false negative), with false implantation (false positive) considered preferable to this. This leads to decision-making optimism, with a subconscious bias for the positive outcome when making subjective decisions (Wang et al., 2021). Additionally, when presented with a decision to grade something in a set, one may do so relative to the set as a whole in a process known as 'value normalization' (Rangel and Clithero, 2012). But this is not the case in biology. If one were to grade several poor embryos in a row, it should not mean that an embryo of slightly better morphological quality should be considered a good embryo. This potential cause of subjectivity (an interesting area where further research is merited) does not exist with the AI approach. An algorithm is trained to grade each embryo independently of other embryos in the cycle. For better or for worse, it is not

burdened with the decisions that have real-world emotional, financial and physical consequences for patients that clinicians are forced to make and which may increase observer false positives when grading embryos or predicting implantation. It is, therefore, reasonable to suggest that AI-based tools could be used in the clinical workflow to help clinicians make critical decisions by combining objective AI assessments with their own experience as well as knowledge of each clinical situation.

As previously mentioned, one limitation of this study is that embryologists are not typically trained to predict implantation or assess likelihood of successful implantation. Instead, they are trained to assess embryo quality using methods that have been found to correlate with implantation or other pregnancy outcomes (Gardner *et al.*, 2000). However, in order to focus on objective solutions, objective outcomes must be selected, and implantation is the outcome with the strongest association with embryo grading (Gardner *et al.*, 2000; Rubio *et al.*, 2012; Thompson *et al.*, 2013). LBR was not selected as the endpoint in an attempt to minimize the confounding factors affecting miscarriage. Another limitation is the decision to use TLI videos as input. TLI is an expensive technology that is currently implemented in a minority of labs (Dolinko *et al.*, 2017; Armstrong *et al.*, 2019). The low adoption rate of TLI limits the application and accessibility of prediction models only capable of interpreting TLI videos, contributing to the motivation behind recent efforts to develop AI-based embryo assessment algorithms for static images (Khosravi *et al.*, 2019; Bormann *et al.*, 2020; VerMilyea *et al.*, 2020; Loewke *et al.*, 2022).

This study was conducted using retrospectively collected data. Future studies should include prospective use of the algorithm to further evaluate its effectiveness in the clinical setting. Continuing with the limitations of the dataset itself, embryos used in this study were almost entirely from ICSI cycles (98%). Although this is an accurate representation of the standard of care at the clinics that supplied the data, future studies should include embryos that were fertilized using conventional IVF. Additionally, despite being blinded to implantation outcomes, participants were aware that all embryos in this study were transferred embryos with KID, which may have biased their results. However, if these limitations were indeed in part responsible for the low PPV and over-estimation of implantation, then it reinforces the notion that subjectivity is a critical factor in the embryologist's performance as the effect of video set bias was minimized due to the balanced nature of the data.

The endpoint of our study was implantation as defined by detection of gestational sac. In similar studies, fetal heartbeat has been used as the endpoint, and can be considered a more robust measure of clinical pregnancy (Tran *et al.*, 2019; Loewke *et al.*, 2022). Unfortunately, records of fetal heartbeat were not available for our datasets. However, as gestational sac identification is often considered the primary criterion for clinical pregnancy/implantation, it is still a clinically relevant endpoint (ESHRE Special Interest Group of Embryology and Alpha Scientists in Reproductive Medicine, 2017; Zegers-Hochschild *et al.*, 2017).

A well-recognized potential limitation of this study is that specific factors used by the model to calculate implantation probability are unidentifiable: the so-called 'black box' approach. Critics suggest that this approach may give rise to a number of unintended complications including preferential selection of embryos with potentially disadvantageous traits, sex selection bias, poor generalization with previously unseen populations and the inability for the embryologist to perform real-time error checking (Afnan *et al.*, 2021). As AI becomes ever more ubiquitous in society, particularly in the medical setting, the issue of explainability is actively being addressed with techniques being developed in order to identify or, at the very least extrapolate contextually, the factors that contribute to an algorithm's output (Arbelaez Ossa *et al.*, 2022).

In summary, the results of this study suggest that embryologists have at least fair agreement when asked to predict implantation from TLI videos of embryos with KID but have far better success when assessing embryos with Gardner scores suggestive of top or poor embryo quality. The fact that a simple DNN model, with off-the-shelf components, was trained with a very small dataset provided by two different clinics, was able to outperform the predictive ability of the embryologist cohort in every measured metric, excluding sensitivity where there was no statistical difference, is very promising. These results suggest the validity of two separate approaches, potentially advocating an assistive focus, especially on moderate morphological embryos: (i) a generalized model trained with a large multicenter dataset containing TLI input from various hardware, potentially capable of being used in any lab as an assistive tool guiding the embryologist; or (ii) a second, more subtle approach, where a prebuilt DNN model is trained on a local scale, essentially accounting for characteristics unique to each clinic. However, some commentators have expressed their doubts for the generalized approach, given the paucity of multicenter validation studies, citing overfitting of the datasets and confounding factors being unidentifiable in the black box approach as possible reasons for poor performance (Curchoe *et al.*, 2020; Afnan *et al.*, 2021). Clinic-tailored models could potentially compensate for confounding variables that differ between locations and clinics and may impact the success of IVF, such as variations in clinical technique, hardware, patient populations, and socio-economic, environmental and healthcare-related factors.

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Data availability

The data underlying this article are available in The Harvard Dataverse at https://doi.org/10.7910/DVN/1Z4HLC.

Acknowledgements

The authors thank Yuval Argaman for his contributions to the preliminary data analyses and discussions, and Avigail Gershenfeld for database organization. They would also like to thank all Embryonics employees for their efforts in this research.

Authors' roles

D.E.F., D.H.S, Y.G.Z., S.R. and A.L.P. were involved in initial study design, with T.A. and D.R. involved in design revision. D.H.S., S.R., Y.G.Z., A.L.P, M.L.L., J.B.N., A.A., and S.M. curated the data. D.E.F.

and A.L.P. were responsible for embryologist project management. D.H.S., S.R, Y.W. and O.P. were responsible for algorithm project management and development. D.E.F., T.A., A.D. and D.R. were responsible for research activity. D.R. and A.M.B. performed the data analysis and visualization. D.E.F. drafted the manuscript. All authors provided critical review and editing of the final draft of the manuscript.

Funding

Embryonics Ltd is an Israel-based company. Funding for the study was partially provided by the Israeli Innovation Authority, grant #74556.

Conflict of interest

D.H.S., Y.G.Z. and A.M.B. are co-founders of Embryonics Ltd. D.E.F., D.R., A.L.P., T.A., Y.W., O.P., A.D. and S.R. are all employees at Embryonics Ltd. M.L.L., J.B.N., A.A. and S.M. have no conflicts of interest to declare.

References

2019 Assisted Reproductive Technology Fertility Clinic and National Summary Report | CDC. 2022. https://www.cdc.gov/art/reports/2019/fertility-clinic.html (4 June 2022, date last accessed).

Afnan MAM, Liu Y, Conitzer V, Rudin C, Mishra A, Savulescu J, Afnan M. Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Hum Reprod Open* 2021;**2021**:hoab040.

Ahlström A, Lundin K, Lind A-K, Gunnarsson K, Westlander G, Park H, Thurin-Kjellberg A, Thorsteinsdottir SA, Einarsson S, Åström M *et al.* A double-blind randomized controlled trial investigating a time-lapse algorithm for selecting Day 5 blastocysts for transfer. *Hum Reprod* 2022;**37**:708–717.

Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology. The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Hum Reprod* 2011;**26**:1270–1283.

Altman DG. *Practical Statistics for Medical Research*. Boca Raton, Florida, USA: Routledge CRC Press, 1990. https://www.routledge.com/Practical-Statistics-for-Medical-Research/Altman/p/book/9780412276309.

Arbelaez Ossa L, Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS. Re-focusing explainability in medicine. *Digit Health* 2022;**8**:205520762210744.

Arce J-C, Ziebe S, Lundin K, Janssens R, Helmgaard L, Sørensen P. Interobserver agreement and intraobserver reproducibility of embryo quality assessments. *Hum Reprod* 2006;**21**:2141–2148.

Armstrong S, Bhide P, Jordan V, Pacey A, Marjoribanks J, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database Syst Rev* 2019;**5**:CD011320.

Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell* 2017;**39**:2481–2495.

Balaban B, Yakin K, Alatas C, Oktem O, Isiklar A, Urman B. Clinical outcome of intracytoplasmic injection of spermatozoa morphologically selected under high magnification: a prospective randomized study. *Reprod Biomed Online* 2011;**22**:472–476.

Barrie A, Homburg R, McDowell G, Brown J, Kingsland C, Troup S. Preliminary investigation of the prevalence and implantation potential of abnormal embryonic phenotypes assessed using time-lapse imaging. *Reprod Biomed Online* 2017;**34**:455–462.

Bendus AEB, Mayer JF, Shipley SK, Catherino WH. Interobserver and intraobserver variation in day 3 embryo grading. *Fertil Steril* 2006;**86**:1608–1615.

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

Bormann CL, Kanakasabapathy MK, Thirumalaraju P, Gupta R, Pooniwala R, Kandula H, Hariton E, Souter I, Dimitriadis I, Ramirez LB *et al.* Performance of a deep learning based neural network in the selection of human blastocysts for implantation. *eLife* 2020;**9**:e55301.

Broer SL, van Disseldorp J, Broeze KA, Dolleman M, Opmeer BC, Bossuyt P, Eijkemans MJ, Mol B-WJ, Broekmans FJ, Broer SL *et al.*; on behalf of the IMPORT study group. Added value of ovarian reserve testing on patient characteristics in the prediction of ovarian response and ongoing pregnancy: an individual patient data approach. *Hum Reprod Update* 2013;**19**:26–36.

Cetinkaya M, Pirkevi C, Yelke H, Colakoglu YK, Atayurt Z, Kahraman S. Relative kinetic expressions defining cleavage synchronicity are better predictors of blastocyst formation and quality than absolute time points. *J Assist Reprod Genet* 2015;**32**:27–35.

Chavez-Badiola A, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Garcia-Sanchez R, Drakeley AJ, Garcia-Sandoval JP. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. *Sci Rep* 2020a;**10**:4394.

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* 2020b;**41**:585–593.

Curchoe CL, Farias AF-S, Mendizabal-Ruiz G, Chavez-Badiola A. Evaluating predictive models in reproductive medicine. *Fertil Steril* 2020;**114**:921–926.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2020. https://openreview.net/forum?id=YicbFdNTTy.

Dolinko AV, Farland LV, Kaser DJ, Missmer SA, Racowsky C. National survey on use of time-lapse imaging systems in IVF laboratories. *J Assist Reprod Genet* 2017;**34**:1167–1172.

ESHRE Special Interest Group of Embryology and Alpha Scientists in Reproductive Medicine. The Vienna consensus: report of an expert meeting on the development of ART laboratory performance indicators. *Reprod Biomed Online* 2017;**35**:494–510.

Farin PW, Britt JH, Shaw DW, Slenning BD. Agreement among evaluators of bovine embryos produced *in vivo* or *in vitro*. *Theriogenology* 1995;**44**:339–349.

Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;**76**:378–382.

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

Gardner DK, Schoolcraft WB, Wagley L, Schlenker T, Stevens J, Hesla J. A prospective randomized trial of blastocyst culture and transfer in *in-vitro* fertilization. *Hum Reprod* 1998;**13**:3434–3440.

Hammond ER, Foong AKM, Rosli N, Morbeck DE. Should we freeze it? Agreement on fate of borderline blastocysts is poor and does not improve with a modified blastocyst grading system. *Hum Reprod* 2020;**35**:1045–1053.

Harton GL, Munné S, Surrey M, Grifo J, Kaplan B, McCulloh DH, Griffin DK, Wells D. Diminished effect of maternal age on implantation after preimplantation genetic diagnosis with array comparative genomic hybridization. *Fertil Steril* 2013;**100**:1695–1703.

Kaser DJ, Racowsky C. Clinical outcomes following selection of human preimplantation embryos with time-lapse monitoring: a systematic review. *Hum Reprod Update* 2014;**20**:617–631.

Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, Sigaras A, Lavery S, Cooper LAD, Hickman C *et al*. Deep learning enables robust assessment and selection of human blastocysts after *in vitro* fertilization. *NPJ Digit Med* 2019;**2**:1–9.

Kingma DP, Ba J. Adam: a method for stochastic optimization. *ArXiv14126980 Cs* [Internet] 2017. http://arxiv.org/abs/1412.6980.

Kirkegaard K, Ahlström A, Ingerslev HJ, Hardarson T. Choosing the best embryo by time lapse versus standard morphology. *Fertil Steril* 2015;**103**:323–332.

Kragh MF, Karstoft H. Embryo selection with artificial intelligence: how to evaluate and compare methods? *J Assist Reprod Genet* 2021;**38**:1675–1689.

Kragh MF, Rimestad J, Berntsen J, Karstoft H. Automatic grading of human blastocysts from time-lapse imaging. *Comput Biol Med* 2019;**115**:103494.

Lambalk CB, Banga FR, Huirne JA, Toftager M, Pinborg A, Homburg R, van der Veen F, van Wely M. GnRH antagonist versus long agonist protocols in IVF: a systematic review and meta-analysis accounting for patient type. *Hum Reprod Update* 2017;**23**:560–579.

Liu Y, Chapple V, Roberts P, Matson P. Prevalence, consequence, and significance of reverse cleavage by human embryos viewed with the use of the embryoscope time-lapse video system. *Fertil Steril* 2014;**102**:1295–1300.e2.

Loewke K, Cho JH, Brumar CD, Maeder-York P, Barash O, Malmsten JE, Zaninovic N, Sakkas D, Miller KA, Levy M *et al*. Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos. *Fertil Steril* 2022;**117**:528–535.

Maheshwari A, McLernon D, Bhattacharya S. Cumulative live birth rate: time for a consensus? *Hum Reprod* 2015;**30**:2703–2707.

Motato Y, de los Santos MJ, Escriba MJ, Ruiz BA, Remohí J, Meseguer M. Morphokinetic analysis and embryonic prediction for blastocyst formation through an integrated time-lapse system. *Fertil Steril* 2016;**105**:376–384.e9.

Munné S. Status of preimplantation genetic testing and embryo selection. *Reprod Biomed Online* 2018;**37**:393–396.

Munné S, Blazek J, Large M, Martinez-Ortiz PA, Nisson H, Liu E, Tarozzi N, Borini A, Becker A, Zhang J *et al*. Detailed investigation into the cytogenetic constitution and pregnancy outcome of replacing mosaic blastocysts detected with the use of high-resolution next-generation sequencing. *Fertil Steril* 2017;**108**:62–71.e8.

Munné S, Kaplan B, Frattarelli JL, Child T, Nakhuda G, Shamma FN, Silverberg K, Kalista T, Handyside AH, Katz-Jaffe M *et al.*; STAR Study Group. Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial. *Fertil Steril* 2019;**112**:1071–1079. e7. Elsevier.

Neyshabur B, Bhojanapalli S, Mcallester D, Srebro N. Exploring generalization in deep learning. *Adv Neural Inf Process Syst* 2017;**30**:5949–5958.

Papanikolaou EG, Camus M, Kolibianakis EM, Van Landuyt L, Van Steirteghem A, Devroey P. *In vitro* fertilization with single blastocyst-stage versus single cleavage-stage embryos. *N Engl J Med* 2006;**354**:1139–1146.

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* 2009;**7**:105.

Petersen BM, Boel M, Montag M, Gardner DK. Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on Day 3. *Hum Reprod* 2016;**31**:2231–2244.

Rangel A, Clithero JA. Value normalization in decision making: theory and evidence. *Curr Opin Neurobiol* 2012;**22**:970–981.

Rocha JC, Passalia F, Matos FD, Maserati MP, Alves MF, de Almeida TG, Cardoso BL, Basso AC, Nogueira MFG. Methods for assessing the quality of mammalian embryos: how far we are from the gold standard? *JBRA Assist Reprod* 2016;**20**:150–158.

Roche K, Racowsky C, Harper J. Utilization of preimplantation genetic testing in the USA. *J Assist Reprod Genet* 2021;**38**:1045–1053.

Rubio I, Kuhlmann R, Agerholm I, Kirk J, Herrero J, Escribá M-J, Bellver J, Meseguer M. Limited implantation success of direct-cleaved human zygotes: a time-lapse study. *Fertil Steril* 2012;**98**:1458–1463.

Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86**:420–428.

Storr A, Venetis C, Cooke S, Kilani S, Ledger W. Time-lapse algorithms and morphological selection of day-5 embryos for transfer: a preclinical validation study. *Fertil Steril* 2018;**109**:276–283. e3.

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**:307–314.

Sundvall L, Ingerslev HJ, Breth Knudsen U, Kirkegaard K. Inter- and intra-observer variability of time-lapse annotations. *Hum Reprod* 2013;**28**:3215–3221.

Tan M, Le QV. EfficientNetV2: smaller models and faster training. *ArXiv210400298 Cs* 2021. http://arxiv.org/abs/2104.00298.

Thompson SM, Onwubalili N, Brown K, Jindal SK, McGovern PG. Blastocyst expansion score and trophectoderm morphology strongly predict successful clinical pregnancy and live birth following elective single embryo blastocyst transfer (eSET): a national study. *J Assist Reprod Genet* 2013;**30**:1577–1581.

Tiegs AW, Tao X, Zhan Y, Whitehead CV, Hanson BM, Kim JG, Osman EK, Seli E, Patounakis G, Gutmann J *et al*. Transfer outcomes of embryos with preimplantation genetic testing for aneuploidy (PGT-A) diagnoses of undetermined reproductive potential:

results from a prospective, blinded, multi-center non-selection study. *Fertil Steril* 2020;**114**:e32.

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**: 1011–1018.

Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science* 1981;**211**:453–458.

Ueno S, Berntsen J, Ito M, Uchiyama K, Okimura T, Yabuuchi A, Kato K. Pregnancy prediction performance of an annotation-free embryo scoring system on the basis of deep learning after single vitrified-warmed blastocyst transfer: a single-center large cohort retrospective study. *Fertil Steril* 2021;**116**:1172–1180.

Van den Abbeel E, Balaban B, Ziebe S, Lundin K, Cuesta MJG, Klein BM, Helmgaard L, Arce J-C. Association between blastocyst morphology and outcome of single-blastocyst transfer. *Reprod Biomed Online* 2013;**27**:353–361.

VerMilyea 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**:770–784.

Walls M, Junk S, Ryan JP, Hart R. IVF versus ICSI for the fertilization of *in-vitro* matured human oocytes. *Reprod Biomed Online* 2012;**25**: 603–607.

Wang S, Zhou C, Zhang D, Chen L, Sun H. A deep learning framework design for automatic blastocyst evaluation with multifocal images. *IEEE Access* 2021;**9**:18927–18934.

van Wely M, Kwan I, Burt AL, Thomas J, Vail A, der Veen FV, Al-Inany HG. Recombinant versus urinary gonadotrophin for ovarian stimulation in assisted reproductive technology cycles. *Cochrane Database Syst Rev* 2011;**2011**:CD005354.

Worrilow KC, Eid S, Woodhouse D, Perloe M, Smith S, Witmyer J, Ivani K, Khoury C, Ball GD, Elliot T *et al.* Use of hyaluronan in the selection of sperm for intracytoplasmic sperm injection (ICSI): significant improvement in clinical outcomes–multicenter, double-blinded and randomized controlled trial. *Hum Reprod* 2013;**28**:306–314.

Zegers-Hochschild F, Adamson GD, Dyer S, Racowsky C, de Mouzon J, Sokol R, Rienzi L, Sunde A, Schmidt L, Cooke ID *et al.* The international glossary on infertility and fertility care, 2017. *Fertil Steril* 2017;**108**:393–406.

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. 2016. https://arxiv.org/abs/1611.03530v2.