






Embryology

BlastAssist: a deep learning pipeline to measure interpretable features of human embryos

Helen Y. Yang ^{1,2,*}, Brian D. Leahy^{1,3}, Won-Dong Jang ⁴, Donglai Wei ⁴, Yael Kalma ⁵, Roni Rahav⁵, Ariella Carmon⁵, Rotem Kopel⁵, Foad Azem⁵, Marta Venturas ¹, Colm P. Kelleher ¹, Liz Cam¹, Hanspeter Pfister ⁴, Daniel J. Needleman^{1,3,†}, and Dalit Ben-Yosef ^{5,6,†}

¹Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA

²Department of Biophysics, Harvard Graduate School of Arts and Sciences, Cambridge, MA, USA

³Department of Applied Physics, Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA

⁴Department of Computer Science, Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA

⁵Department of Reproduction and IVF, Lis Maternity Hospital Tel-Aviv Sourasky Medical Center, Tel Aviv, Israel

⁶Department of Cell and Developmental Biology, Sackler Faculty of Medicine, Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv, Israel

*Correspondence address. Needleman Lab, Department of Molecular and Cellular Biology, Harvard University, 52 Oxford St, Room 359.10, Cambridge, MA 02138, USA. Telephone: +1 617.384.6739; E-mail: helen_yang@fas.harvard.edu  <https://orcid.org/0000-0001-6257-266x>

[†]These authors contributed equally as co-last authors.

ABSTRACT

STUDY QUESTION: Can the BlastAssist deep learning pipeline perform comparably to or outperform human experts and embryologists at measuring interpretable, clinically relevant features of human embryos in IVF?

SUMMARY ANSWER: The BlastAssist pipeline can measure a comprehensive set of interpretable features of human embryos and either outperform or perform comparably to embryologists and human experts in measuring these features,

WHAT IS KNOWN ALREADY: Some studies have applied deep learning and developed 'black-box' algorithms to predict embryo viability directly from microscope images and videos but these lack interpretability and generalizability. Other studies have developed deep learning networks to measure individual features of embryos but fail to conduct careful comparisons to embryologists' performance, which are fundamental to demonstrate the network's effectiveness.

STUDY DESIGN, SIZE, DURATION: We applied the BlastAssist pipeline to 67 043 973 images (32 939 embryos) recorded in the IVF lab from 2012 to 2017 in Tel Aviv Sourasky Medical Center. We first compared the pipeline measurements of individual images/embryos to manual measurements by human experts for sets of features, including: (i) fertilization status ($n = 207$ embryos), (ii) cell symmetry ($n = 109$ embryos), (iii) degree of fragmentation ($n = 6664$ images), and (iv) developmental timing ($n = 21 036$ images). We then conducted detailed comparisons between pipeline outputs and annotations made by embryologists during routine treatments for features, including: (i) fertilization status ($n = 18 922$ embryos), (ii) pronuclei (PN) fade time ($n = 13 781$ embryos), (iii) degree of fragmentation on Day 2 ($n = 11 582$ embryos), and (iv) time of blastulation ($n = 3266$ embryos). In addition, we compared the pipeline outputs to the implantation results of 723 single embryo transfer (SET) cycles, and to the live birth results of 3421 embryos transferred in 1801 cycles.

PARTICIPANTS/MATERIALS, SETTING, METHODS: In addition to EmbryoScopeTM image data, manual embryo grading and annotations, and electronic health record (EHR) data on treatment outcomes were also included. We integrated the deep learning networks we developed for individual features to construct the BlastAssist pipeline. Pearson's χ^2 test was used to evaluate the statistical independence of individual features and implantation success. Bayesian statistics was used to evaluate the association of the probability of an embryo resulting in live birth to BlastAssist inputs.

MAIN RESULTS AND THE ROLE OF CHANCE: The BlastAssist pipeline integrates five deep learning networks and measures comprehensive, interpretable, and quantitative features in clinical IVF. The pipeline performs similarly or better than manual measurements. For fertilization status, the network performs with very good parameters of specificity and sensitivity (area under the receiver operating characteristics (AUROC) 0.84–0.94). For symmetry score, the pipeline performs comparably to the human expert at both 2-cell ($r = 0.71 \pm 0.06$) and 4-cell stages ($r = 0.77 \pm 0.07$). For degree of fragmentation, the pipeline (acc = 69.4%) slightly under-performs compared to human experts (acc = 73.8%). For developmental timing, the pipeline (acc = 90.0%) performs similarly to human experts (acc = 91.4%). There is also strong agreement between pipeline outputs and annotations made by embryologists during routine treatments. For fertilization status, the pipeline and embryologists strongly agree (acc = 79.6%), and there is strong correlation between the two measurements ($r = 0.683$). For degree of fragmentation, the pipeline and embryologists mostly agree (acc = 55.4%), and there is also strong correlation between the two measurements ($r = 0.648$). For both PN fade time ($r = 0.787$) and time of blastulation ($r = 0.887$), there's strong correlation between the pipeline and embryologists. For SET cycles, 2-cell time ($P < 0.01$) and 2-cell symmetry ($P < 0.03$) are significantly correlated with implantation success rate, while other features showed correlations with implantation success without statistical significance. In addition, 2-cell time ($P < 5 \times 10^{-11}$), PN fade time ($P < 5 \times 10^{-10}$), degree of fragmentation on

Received: May 18, 2023. Revised: January 5, 2024. Editorial decision: January 30, 2024.

© The Author(s) 2024. 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

Day 3 ($P < 5 \times 10^{-4}$), and 2-cell symmetry ($P < 5 \times 10^{-3}$) showed statistically significant correlation with the probability of the transferred embryo resulting in live birth.

LIMITATIONS, REASONS FOR CAUTION: We have not tested the BlastAssist pipeline on data from other clinics or other time-lapse microscopy (TLM) systems. The association study we conducted with live birth results do not take into account confounding variables, which will be necessary to construct an embryo selection algorithm. Randomized controlled trials (RCT) will be necessary to determine whether the pipeline can improve success rates in clinical IVF.

WIDER IMPLICATIONS OF THE FINDINGS: BlastAssist provides a comprehensive and holistic means of evaluating human embryos. Instead of using a black-box algorithm, BlastAssist outputs meaningful measurements of embryos that can be interpreted and corroborated by embryologists, which is crucial in clinical decision making. Furthermore, the unprecedentedly large dataset generated by BlastAssist measurements can be used as a powerful resource for further research in human embryology and IVF.

STUDY FUNDING/COMPETING INTEREST(S): This work was supported by Harvard Quantitative Biology Initiative, the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (award number 1764269), the National Institute of Health (award number R01HD104969), the Perelson Fund, and the Sagol fund for embryos and stem cells as part of the Sagol Network. The authors declare no competing interests.

TRIAL REGISTRATION NUMBER: Not applicable.

Keywords: IVF / EmbryoScope / deep learning / computer vision / embryo grading / machine learning / interpretable features

Introduction

IVF has revolutionized the treatment of human infertility and more than 3 million IVF cycles are now performed each year worldwide (Adamson et al., 2022). Despite the ubiquity of IVF treatments, their success rate remains relatively low: only ~35% of cycles result in live birth in the USA (CDC, 2021), leading to high financial and emotional costs.

While increasing the number of embryos to transfer increases the potential for live births, it also increases the risks for multiple pregnancies with associated maternal and offspring morbidity and mortality (Norwitz et al., 2005). Therefore, clinical standards strongly recommend single embryo transfer (SET) (Lee et al., 2016), which greatly increases the need to accurately evaluate and select embryos.

Manual morphological grading is still the most widely used method for embryo evaluation (Ebner et al., 2003; Mastenbroek et al., 2011). Pre-implantation Genetic Testing for Aneuploidies (PGT-A) is increasingly used for embryo selection, though it is invasive and its efficacy is controversial (Mastenbroek et al., 2008; Lee et al., 2015; Paulson, 2020). Time-lapse microscopy (TLM) incubators have been adopted in many clinics to culture embryos while collecting continuous movies of pre-implantation development (Kirkegaard et al., 2012; Dolinko et al., 2017; Armstrong et al., 2019). TLM systems provide significantly more information than traditional manual morphological grading (Campbell et al., 2013, 2014; Amir et al., 2019). Many clinics that utilize TLM systems rely on manual evaluation of embryo movies, which is extremely time-consuming and subjective. In a recent randomized controlled trial (RCT), the current method of utilizing TLM failed to improve IVF outcomes relative to manual morphological grading alone (Ahlström et al., 2022), which might be due to not all the information from TLM being leveraged. An automated and comprehensive means of extracting clinically and biologically relevant information from TLM movies would be greatly beneficial to embryologists and has the potential to improve IVF outcomes.

Machine learning (ML) algorithms, which are based on constructing predictive models from the (often subtle) associations present in ‘training’ data, have been highly successful at analyzing large and complex datasets. Compared to humans, ML algorithms can consistently, rapidly, and accurately process large amounts of data at very low cost (Jordan and Mitchell, 2015), making them a promising means for aiding in embryo evaluation through TLM movies. There have been several prior attempts to develop ML algorithms for use in embryo evaluation and selection. Some previous ML algorithms were developed to directly

link images or movies of embryos to the probability of an embryo implanting (Bormann et al., 2020; Chavez-Badiola et al., 2020; Silver et al., 2020) and developing to fetal heartbeats (Tran et al., 2019; VerMilyea et al., 2020). However, such ‘black-box’ approaches have a number of disadvantages (Rudin, 2019; Afnan et al., 2021a,b): (i) The lack of implantation ground truth in most IVF cycles makes the training data, and hence the resulting algorithms, biased towards SET cycles, which might not be always representative; (ii) there are numerous confounders, including patient age, BMI, uterine receptivity, sperm quality, culture conditions, and variations in treatment procedures, which may make ML algorithms trained solely on embryo images/movies not generalizable to different patient populations; and (iii) most fundamentally, typical ML algorithms are designed to aid inference, but the process of selecting the best embryo for transfer entails a causal question—what will the result of the IVF treatment be if we transfer this particular embryo?—rather than an inferential question—given that this embryo was transferred, what is the probability that it implanted?

ML algorithms that extract biologically and clinically relevant features from TLM movies have the potential to overcome the limitations of black-box approaches. Automatically measuring such features would aid embryologists in their current practices, and could be combined with interpretable, statistical models of oocyte and embryo development (Leahy et al., 2021) to guide embryo selection. A number of prior studies have developed ML algorithms to measure interpretable features, but, so far, these have been mostly limited to simple, individual features such as cell stage up to 5- or 8-cell (Khan et al., 2016; Lau et al., 2019; Malmsten et al., 2019), blastocyst segmentation of inner cell mass (ICM) and trophectoderm (TE) (Rad et al., 2018, 2020; Harun et al., 2019), and blastocyst grading (Khosravi et al., 2019; Kragh et al., 2019). In addition, automated measurements might be timesaving and assist embryologists, but to be useful, their accuracy must be at least comparable to that of embryologists. To the best of our knowledge, no prior studies have systematically compared automated measurements to those of human experts and embryologists (Simopoulou et al., 2018; Sfakianoudis et al., 2022).

Here, we built off our prior works (Jang et al., 2023; Leahy et al., 2020; Lukyanenko et al., 2021) and developed BlastAssist, a holistic pipeline to measure a comprehensive set of interpretable features that are clinically and biologically relevant, including: (i) fertilization status, (ii) cell symmetry, (iii) degree of fragmentation, (iv) developmental timing, and (v) size of the ICM and TE

and the dynamics of blastocyst expansion. We used BlastAssist to analyze movies of 32 939 human embryos (67 043 973 images), resulting in an unprecedented dataset that will be a powerful resource for studying human pre-implantation embryology. To characterize the accuracy of the BlastAssist pipeline, we conducted detailed comparisons between the pipeline outputs to four different metrics: (i) manual annotations from a panel of human experts in idealized situations, (ii) annotations previously made by embryologists during the course of routine IVF treatments, (iii) implantation success rate of SET cycles, and (iv) the likelihood that the transferred embryo would result in live birth.

Materials and methods

Study design and dataset

The EmbryoScope™ dataset was collected from the IVF unit of Sourasky Medical Center in Tel Aviv, Israel from treatment cycles performed between 2012 and 2017. EmbryoScope™ is the most widely used TLM system for IVF (Dolinko et al., 2017). It utilizes Hoffman modulation contrast (HMC) microscopy (Hoffman and Gross, 1975). The dataset consists of 32 939 embryos, imaged every 20 min at seven different focal planes up to the first 5 days of development, yielding 67 043 973 JPEG files, each 500 × 500 pixels in size. Along with the image data, standard clinical annotations were also recorded for these embryos. The clinical annotations include: (i) electronic health record (EHR) data, (ii) treatment information, such as fertilization method, hormone dosage, and culture media, (iii) embryo grading and annotations, such as developmental timing, degree of fragmentation on Days 2 and 3, PN count, and fade time, and (iv) treatment outcomes, such as beta-HCG (b-HCG) and live births. These clinical annotations were used to compare clinic and network measurements at four different levels: Comparison of BlastAssist measurements to: (i) measurements performed by human expert(s) in ideal situations, (ii) clinical measurements during routine treatments, (iii) implantation outcomes for SET cycles, and (iv) live birth results for cycles where up to four embryos were transferred.

Metrics

In addition to metrics commonly used in IVF clinics, we implemented quantitative metrics to better represent the results of BlastAssist.

We define the symmetry score as a function of time $S(t)$ as the standard deviation of the areas normalized by their mean. We calculate the embryo's symmetry score as the time-averaged symmetry for the entire 2-cell or 4-cell stage.

We define the average thickness of TE and zona pellucida (ZP) as the average of the closest distances from each point on the medial axis (Breu et al., 1995) of the TE or ZP to the boundary multiplied by 2.

When ICMs first form, their sizes increase significantly at first and then fluctuate around a constant value. In addition, ICMs are known to have significant movement throughout the blastocyst stage, which means that the ICMs can move in and out of the focal plane during imaging, which can dramatically change the ICM's detected size. We define the average ICM size as the average area of detected ICM over time if the ICM size is relatively constant for approximately 5 h or more.

We define blastocyst expansion rate as the slope of the linear regression fit of the size of the blastocyst diameter over time if the Pearson correlation coefficient is larger than 0.85, to eliminate any blastocysts that do not have a steady growth rate.

Statistics

Pearson's correlation coefficient (r) (Freedman et al., 2007; Benesty et al., 2008) was used to evaluate associations between two measurements.

For implantation results of the 723 SET cycles, the cycles are divided into two to four classes for each feature and the success rates for implantation are calculated. For continuous features, the classes are divided based on both the biological meaning of the feature and the data distribution. The error bars are calculated $\sigma = \sqrt{P(1-P)/N}$, where P is the probability which is the implantation success rate and N is the sample size. Pearson's χ^2 test (McHugh, 2013) was performed on each feature to evaluate the statistical independence of the feature and the implantation results.

For live birth results of the 3421 embryos (1801 cycles) where up to four embryos were transferred, we used a previously developed method (Leahy et al., 2021) to estimate the probability of an implanted embryo resulting in live birth as a function of individual features. We estimate this probability with a Bayesian approach (see Supplementary Data File S1: Correlations to live birth). Two-sided t-test (Student, 1908) was performed on each feature to evaluate the statistical significance of the slope of the fitting.

Ethics

The study was approved by the Internal Review Board (IRB) of both Tel Aviv Sourasky Medical Center (IRB 606/17) and Harvard University (IRB 18-0532).

Results

BlastAssist pipeline

In our previous works (Jang et al., 2023; Leahy et al., 2020; Lukyanenko et al., 2021), we trained five individual neural networks: (i) ZP segmentation, (ii) developmental stage classifier, (iii) degree of fragmentation classifier, (iv) PN detector, and (v) blastomere detector to measure comprehensive metrics during pre-implantation development. In the current study we trained an additional network: (vi) blastocyst segmentation (see Data availability and Supplementary Data File S1) (Supplementary Tables S1 and S2).

By integrating these six networks into a holistic pipeline, which we call BlastAssist, we have created a means to perform automated measurements for each embryo, throughout all developmental stages during pre-implantation development (Fig. 1, Supplementary Figs S1, S2, S3, and S4, Video 1, Supplementary Videos S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, and S11). We applied the BlastAssist pipeline to a dataset of 67 043 973 images (32 939 time-lapse movies of embryos), resulting in an unprecedentedly large human embryo dataset containing comprehensive and quantitative measurements (see Supplementary Table S3 for patient demographic and cycle information).

Comparison of BlastAssist pipeline to human experts

In general, computer vision (CV) networks are evaluated based on their performances of various metrics on the test set, which we have performed in our previous study (Leahy et al., 2020). Another common evaluation for the efficacy of CV networks involves letting multiple human labelers or experts in the field perform the same measurements or tasks as the networks. This type of comparison is usually conducted because for CV networks to be effective, they have to perform similarly to or

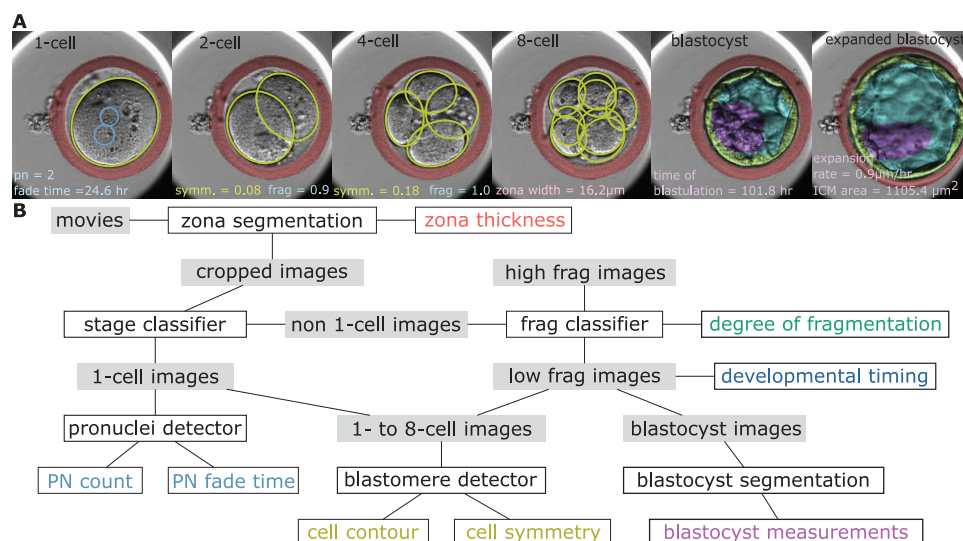
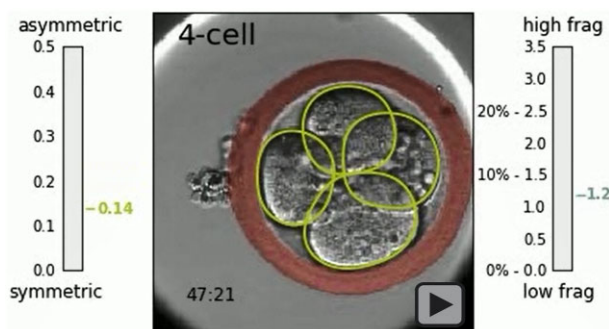


Figure 1. Overview of the BlastAssist pipeline of networks. (A) An example of BlastAssist outputs on a time-lapse video of an embryo (see [Video 1](#)) with measurements highlighted. Note: the stage classifier currently does not differentiate between blastocyst and expanded blastocyst, the label of 'expanded blastocyst' is manually added. (B) Here, we present the holistic pipeline of our networks with all the biological and clinical measurements. We first evaluate all the movies with the zona pellucida (zona) segmentation and crop the images to the embryo. We measure the zona thickness from the result. We then evaluate all the cropped images with the stage classifier and images that are determined to be 1-cell will be evaluated by the pronuclei detector, where we obtain pronuclei (PN) count and fade time. All other images will be evaluated by the fragmentation (frag) classifier, where we obtain the degree of fragmentation. Only the images with low fragmentation (<1.5) will be evaluated further. Developmental timing measurements are obtained from these images with low fragmentation. Images that are identified by the stage classifier to be 1- to 8-cell images will be evaluated by the blastomere detector where cell contour and symmetry are measured. Blastocyst images will be segmented to obtain blastocyst measurements.



Video 1. Sample video of BlastAssist pipeline applied to a human embryo. Full video of [Fig. 1A](#). A 2 pronuclei (PN) embryo that was produced by intracytoplasmic sperm injection (ICSI) and developed to a cleavage stage embryo with high cell symmetry and low fragmentation, started blastulation at 101-h post-fertilization and developed to a high-quality expanded blastocyst.

outperform human experts, depending on the variable we are trying to measure. These human expert measurements are usually performed under idealized and controlled circumstances. In this study, we not only performed this type of comparison between human experts and the BlastAssist pipeline, we also performed comparison between clinical results and the pipeline, providing a comprehensive evaluation of the efficacy of the pipeline, from both a CV and clinical perspective, which most studies fail to do.

For each network in the pipeline, depending on the complexity of the specific task and the estimated human accuracy, we have either one or a few human experts perform the same tasks by hand. For simple tasks where we expect relatively high human accuracies, such as counting number of PN, we take the human expert input as the ground truth dataset. For complicated tasks

where we expect disagreements between human experts, such as labeling the degree of fragmentation, we have multiple human experts perform the same task, and create a ground-truth dataset using the majority consensus of these labelers and the network. We compared the accuracy of the network and of the human experts to the ground truth dataset.

For fertilization status, the network performs with very good parameters of specificity and sensitivity (areas under the receiver operating characteristic (AUROC) curve of 0.84–0.94) ([Fig. 2A](#)). We also measured the human expert's true positive and false positive rate of labeling embryos as having 0, 1, 2, or ≥3PN. The results show that the network performs as well as or better than the human expert at identifying embryos with either 0PN or 1PN, and it performs slightly less well than the human expert at identifying whether an embryo has 2PN or ≥3PN.

For cell symmetry, we compared the network measurement to the cell symmetry scores generated from a human expert's manually traced cell boundaries for each image. Symmetry scoring of the network and of the human expert are highly correlated for both 2-cell ($r=0.71\pm 0.06$) and 4-cell ($r=0.77\pm 0.07$) embryos ([Fig. 2B](#)). The results show that the network is highly accurate compared to the human expert, and the accuracy is consistent with both symmetric and asymmetric embryos.

Degree of fragmentation is an important parameter in determining embryo quality, but it is considered to be relatively subjective and we see greater variability between embryologists in determining the degree of fragmentation ([Paternot et al., 2011](#)). Therefore, for degree of fragmentation, five human experts labeled the same test set, where each labeler annotated the degree of fragmentation for every image. We then created a ground-truth label for each image using the majority vote of these five labelers plus network, excluding images that do not have a majority consensus. We compared the network to the ground truth dataset, and the network's overall accuracy of classifying the

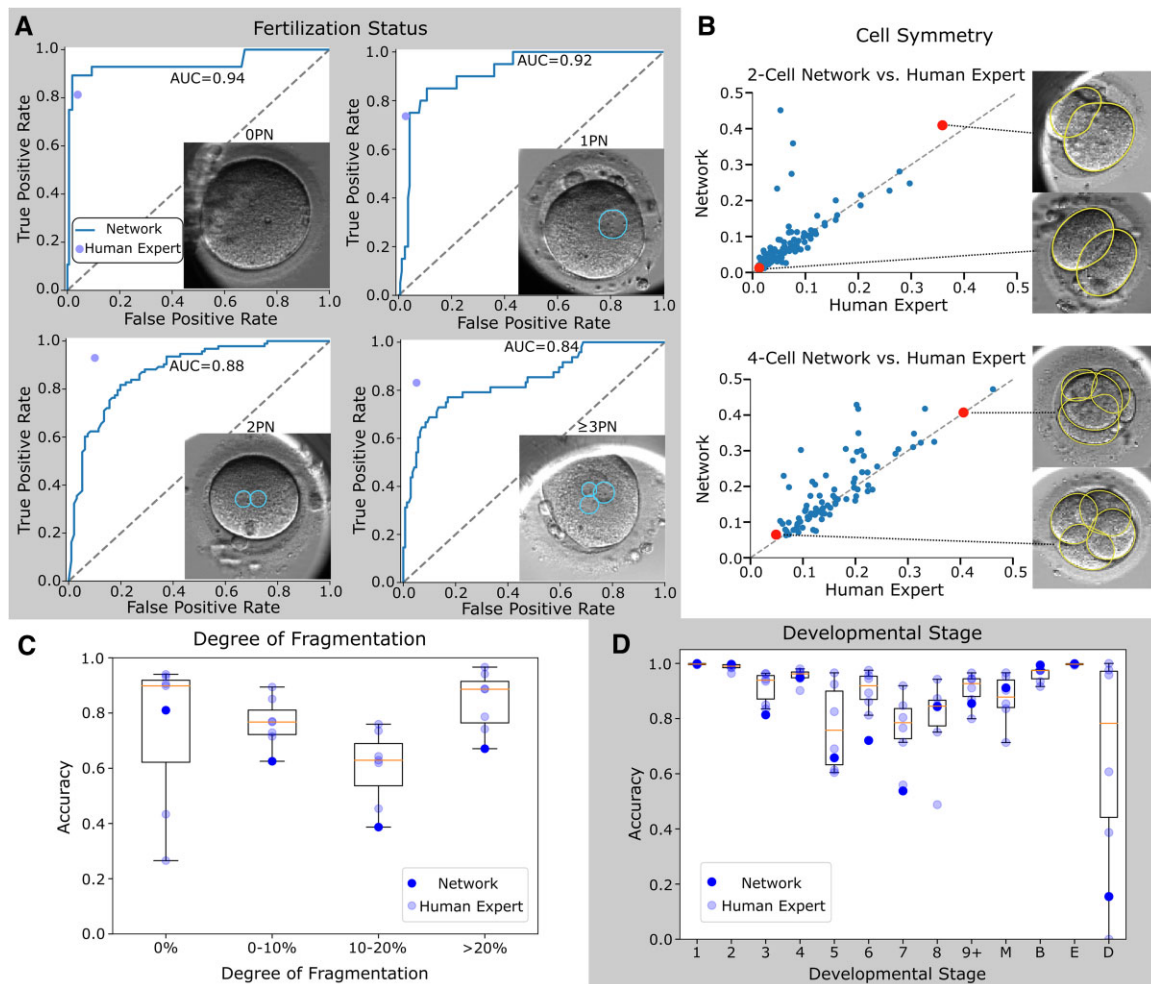


Figure 2. Comparison of BlastAssist outputs to human experts. (A) ROC curves of network performance in blue on OPN (pronuclei) (upper left), 1PN (upper right), 2PN (lower left), and ≥ 3 PN (lower right) embryos with comparison to human expert performance in light blue dots. Pronuclei detections are highlighted in blue in sample images. $n = 207$ embryos. (B) Comparison of network (y-axis) to human expert measurements (x-axis) of 2- (upper) and 4-cell (lower) stage symmetry score. The red dots corresponds to sample images of symmetric and asymmetric embryos in both stages, with cell contour detection highlighted in yellow. $n = 109$ embryos. (C) Comparison of network (dark blue dots) and five human experts (light blue dots) performance on different degrees of fragmentation. $n = 6664$ images. (D) Comparison of network (dark blue dots) and five human experts (light blue dots) performance on different developmental stages. $n = 21\,036$ images. The x-axis labels correspond to 1– to 9+ cell, morula (M), blastocyst (B), empty well (E), and degenerate (D).

degree of fragmentation is 69.4%. We also compared the assessment of each human expert to the ground-truth dataset; the five human experts' overall accuracies were relatively high (76.5%, 79.4%, 77.6%, 66.4%, and 69.2%), with an average of 73.8%. The network performs comparably to human experts at identifying embryos with 0% fragmentation, and it slightly underperforms relative to human experts at identifying embryos with 0–10%, 10–20%, and >20% fragmentation (Fig. 2C).

For developmental stage, five human experts each labeled the test set, where each labeler annotated the developmental stage for every image. We created a ground-truth label for each image using the majority consensus of these labelers and the network. We compared the accuracy of the network and of the human experts to this ground truth dataset (Fig. 2D). Overall, the network has a 90.0% accuracy. The five human experts' overall accuracies were as high as 93.2%, 92.9%, 92.4%, 91.8%, and 86.8%, with an average of 91.4%. The network tends to perform with high accuracy on the physiologically typical stages of 2-, 4-, 8-cells and relatively lower accuracy on the transition stages of 3-, 5-, 6-, and 7-cells. This is expected because the transition stages

are brief, and thus less common, resulting in fewer images in these stages for training, and the number of cells is not always well-defined during cell divisions. The network is very accurate at identifying morula, blastocysts, and empty wells. For degenerate embryos, the occurrences are small but being able to deselect them is very important. Since it's very difficult to distinguish between highly fragmented embryos and degenerated embryos, the results varies greatly amongst human experts.

Comparison of BlastAssist pipeline to clinical measurements

Since the human expert performance is from an idealized situation where every image is annotated in detail, it can differ from embryologists' performance in an actual clinical setting, where embryo annotations are performed only at distinct time points, and with significantly more time constraints. Therefore, we further evaluated the BlastAssist performance by comparing the pipeline measurements to the available clinical annotations recorded during routine treatments.

First, we present the general description of the clinical variables measured by the BlastAssist pipeline (Supplementary Fig. S5) and comparison to clinical annotations when available (Table 1). The general distributions of BlastAssist and clinical measurements are very similar, with discrepancies in PN count and degree of fragmentation which we address later in this section.

There are four measurements that are representative and quantitative that BlastAssist evaluates which were also previously recorded by embryologists in the IVF lab during the course of routine treatments: PN count, PN fade time, degree of fragmentation, and the time of blastulation. We compared the BlastAssist measurements to available clinical data from our dataset of 32 939 embryos, including both side-by-side comparison of the data distribution and per-embryo comparison of each measurement (Fig. 3).

For PN count, the network and embryologists strongly agree, with strong correlation of between the two measurements ($\text{acc} = 79.6\%$, $r = 0.683$). They mostly agree on 1PN and 2PN embryos, with small disagreements on 0PN and ≥ 3 PN embryos (Fig. 3A). The small discrepancies on number of PN mostly come from these following cases: (i) an air bubble or other impurities obstructing the view of the microscope, (ii) mis-identifying 2PN with vacuole as ≥ 3 PN, (iii) mis-identifying 2PN as 1PN when one of the pronuclei is extremely faint and/or faded after very few frames. In these cases, the classifier cannot function properly and can misidentify the number of PN. However, these cases can be easily spot-checked and verified by an embryologist. For PN fade time, the network and embryologists mostly agree, and their measurements have almost identical distributions with a strong correlation ($r = 0.787$) (Fig. 3B). For degree of fragmentation, the network and embryologists agree 55.4% of the time, with most of the disagreement coming from two adjacent classes. This discrepancy may be coming from the low precision of human labeling of this feature. However, there is still a strong correlation between the two measurements ($r = 0.648$) (Fig. 3C). For time of blastulation, the network and embryologists mostly agree, and their measurements have almost identical distributions and a strong correlation ($r = 0.887$) (Fig. 3D).

Since the clinical annotations consist of single measurements for each embryo, it is not possible to use this data to quantify uncertainties or determine a ground truth. Nevertheless, the close agreement between the clinical annotations and the BlastAssist outputs suggests that BlastAssist performs comparably to embryologists in a clinical setting.

Associations of BlastAssist outputs to clinical implantation results

In addition to comparison to human expert or clinical annotations in idealized situations and during routine treatments, we also compared selected BlastAssist results to IVF outcomes. Association with clinical outcomes can provide insight into how well the pipeline might be able to predict IVF outcomes once implemented into clinical environments. In this study, we evaluated associations to clinical outcomes in two different ways. First, we used 723 SET cycles with known implantation results for every embryo transferred (Fig. 4). SET cycles allow us to have the most accurate information on the implantation success rate of individual embryos.

We evaluated the association of SET cycle implantation results to clinical features that BlastAssist measures. Here implantation success is defined as positive (>25 IU/L) b-HCG results. There are five main clinical features that show strong correlations with implantation success. For Day 2 degree of fragmentation, embryos with higher degree of fragmentation have lower implantation success rates ($P < 0.3$) (Fig. 4A). For developmental timing, delayed cleavage into 2-cell stage is negatively correlated with implantation success rate, with embryos undergoing the first cleavage at <24 h presenting the highest chance for implantation ($P < 0.01$) (Fig. 4B). For 2-cell symmetry, embryos with higher symmetry (less than 0.2) (see Materials and methods: Metrics) have higher implantation success rate ($P < 0.03$) (Fig. 4C). When analyzing blastocysts, our results show that embryos with higher blastocyst expansion rate (>1.5 $\mu\text{m}/\text{h}$) have higher implantation success rate ($P < 0.2$) (Fig. 4D). ICM size of blastocysts also shows strong correlation with implantation success rate, with blastocysts presenting ICM size over 1000 μm^2 demonstrating higher implantation rates ($P < 0.5$) (Fig. 4E). Amongst the features, developmental timing and cell symmetry are significantly

Table 1. Distributions of EmbryoScope™ dataset measured by BlastAssist and by clinical annotations.

	The entire dataset (N = 32 939 embryos)	Dataset where clinical annotations are available (N = 12 737 embryos)	
	BlastAssist	BlastAssist	Clinical annotation
PN count			
2PN (normal)	63.3%	79.2%	94.7%
0PN (unfertilized)	16.2%	4.4%	1.7%
1PN (abnormal)	13.6%	10.9%	3.3%
≥ 3 PN (abnormal)	6.9%	5.4%	0.3%
PN fade time (h)	33.9 ± 23.7	27.1 ± 11.7	25.6 ± 4.1
1-cell ZP thickness (μm)	19.6 ± 2.7	19.5 ± 2.7	–
2-cell time (h)	30.8 ± 13.0	28.6 ± 7.2	28.7 ± 5.3
4-cell time (h)	42.4 ± 12.7	41.2 ± 9.4	41.2 ± 6.6
Symmetry score			
2-cell	0.12 ± 0.10	0.11 ± 0.09	–
4-cell	0.20 ± 0.09	0.19 ± 0.08	–
Day 2 degree of fragmentation			
0%	54.9%	46.7%	29.0%
0–10%	30.5%	36.9%	43.9%
10–20%	9.0%	10.2%	20.0%
$>20\%$	5.7%	6.2%	7.0%
Blastulation time (h)	101.2 ± 15.5	102.6 ± 12.9	101.9 ± 8.3
ICM size (μm^2)	1017.6 ± 515.8	999.3 ± 482.0	–
Blastocyst expansion rate ($\mu\text{m}/\text{h}$)	1.6 ± 0.7	1.7 ± 0.7	–

Data are presented as mean ± SD or percentage. ICM, inner cell mass; PN, pronuclei; ZP, zona pellucida.

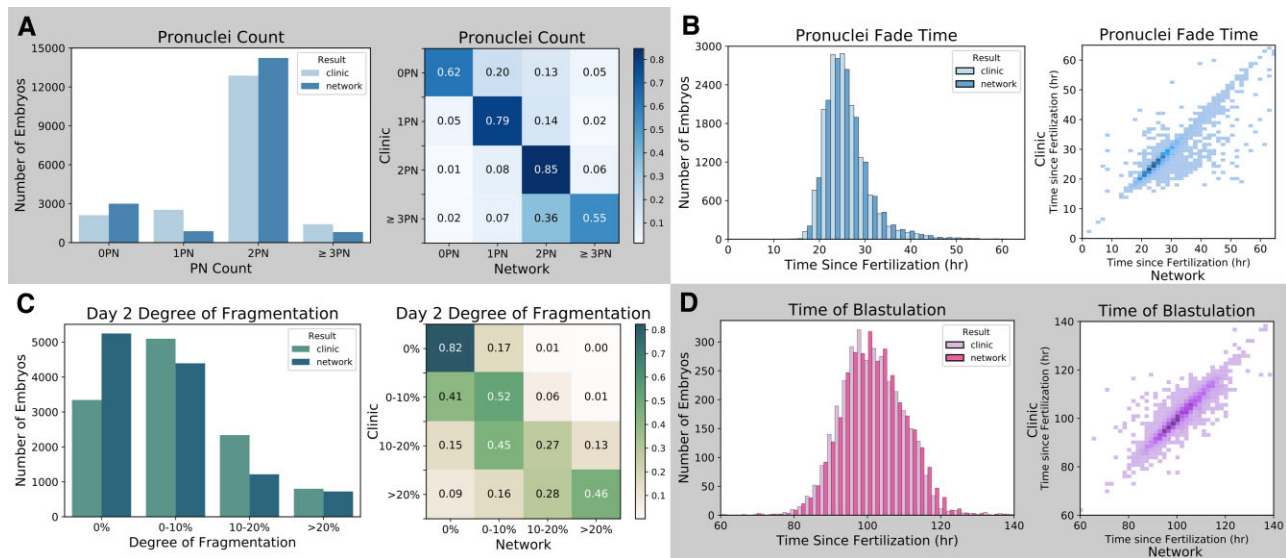


Figure 3. Comparison of BlastAssist outputs to clinical annotations. (A) Comparison of pronuclei (PN) count measurements between clinic and network, with overall number of embryos per class (left) and confusion matrix of measurements per embryo (right). $n = 18922$ embryos. Each intersection of the confusion matrix represents the amount of embryos identified as one class by the network and as the corresponding class by the clinic. The values are normalized between 0 and 1 based on the clinical measurements. (B) Comparison of PN fade time measurements between clinic and network, with overall number of embryos per class (left) and scatterplot of measurements per embryo (right). $n = 13781$ embryos. (C) Comparison of Day 2 degree of fragmentation measurements between clinic and network, with overall number of embryos per class (left) and confusion matrix of measurements per embryo (right). $n = 11582$ embryos. Each intersection of the confusion matrix represents the amount of embryos identified as one class by the network and as the corresponding class by the clinic. The values are normalized between 0 and 1 based on the clinical measurements. (D) Comparison of time of blastulation measurements between clinic and network, with overall number of embryos per class (left) and scatterplot of measurements per embryo (right). $n = 3266$ embryos.

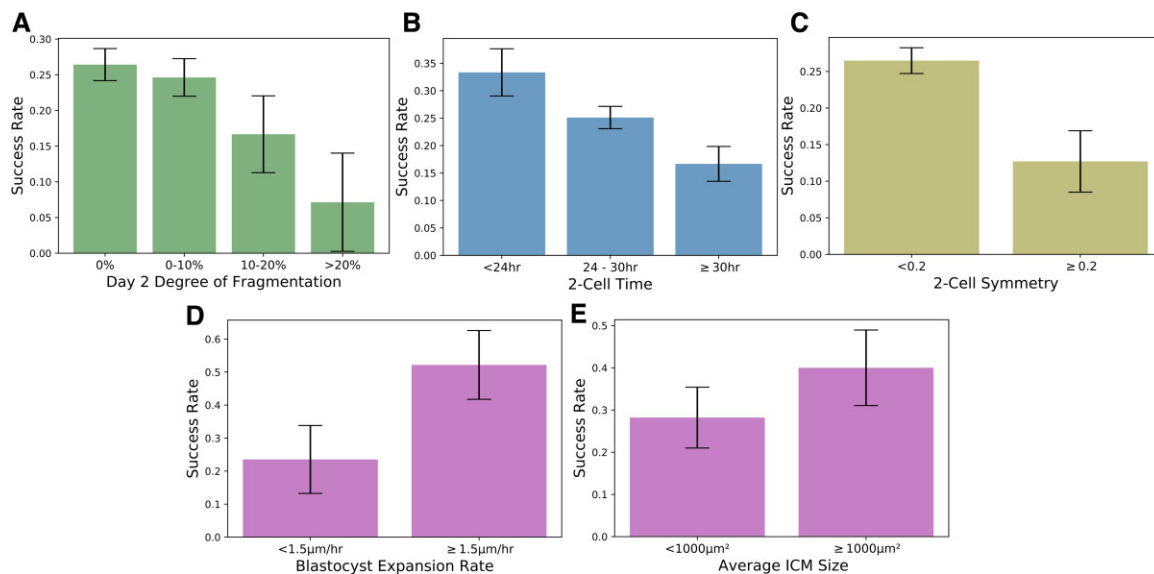


Figure 4. Associations of BlastAssist outputs to clinical implantation results. Seven hundred and twenty-three single embryo transfer (SET) cycles were included. (A) Implantation success rate for embryos with 0% ($n = 386$), 0–10% ($n = 268$), 10–20% ($n = 48$), and >20% ($n = 14$) fragmentation on Day 2 post-fertilization. $\chi^2 = 4.60$, $P < 0.3$. (B) Implantation success rate for embryos that developed to 2-cell stage in <24 h ($n = 120$), 24–30 h ($n = 454$), and ≥ 30 h ($n = 138$) post-fertilization. $\chi^2 = 9.59$, $P < 0.01$. (C) Implantation success rate for embryos with <0.2 ($n = 627$), and ≥ 0.2 ($n = 63$) cell symmetry at 2-cell stage. $\chi^2 = 5.05$, $P < 0.03$. (D) Implantation success rate for embryos that expand at <1.5 $\mu\text{m}/\text{h}$ ($n = 17$), and ≥ 1.5 $\mu\text{m}/\text{h}$ ($n = 23$) during blastocyst stage. $\chi^2 = 2.25$, $P < 0.2$. (E) Implantation success rate for embryos with inner cell mass (ICM) size <1000 μm^2 ($n = 39$), and ≥ 1000 μm^2 ($n = 30$) during blastocyst stage. $\chi^2 = 0.60$, $P < 0.5$.

correlated with implantation success rate ($P < 0.05$). Other features show correlations with implantation success without statistical significance, likely due to the small sample sizes. For example, very few embryos with higher than 20% fragmentation are transferred as singletons in typical IVF cycles, therefore only a small amount of those can be included in the analysis ($n = 14$).

Associations of BlastAssist outputs to clinical live birth results

SET cycles give us the most accurate information on individual embryos' clinical outcome. However, SET cycles have usually been limited to younger patients and patients with good prognosis (De Neubourg and Gerris, 2003; van Montfort et al., 2005),

resulting in a relatively small sample size, even from our large clinical dataset. Therefore, we used a previously developed Bayesian inference approach (Leahy et al., 2021) to investigate associations of BlastAssist measurements to embryos in cycles where up to four embryos were transferred (see [Supplementary Data File S1: Correlations to live birth](#)). Here we compared selected BlastAssist results to the clinical live births outcomes of 3421 embryos that were transferred (1801 cycles) (Fig. 5). We found the correlations between the probability of an embryo resulting in a live birth and individual BlastAssist measurements (Fig. 5 and [Supplementary Fig. S6](#)).

Delayed PN fade time is strongly negatively correlated to the probability of an embryo resulting in live birth ($r = -0.50 \pm 0.08$, $t = -6.25$, $P < 5 \times 10^{-10}$) (Fig. 5A). ZP thickness during the 1-cell stage shows no significant correlation to the probability of an embryo resulting in live birth ($r = 0.04 \pm 0.06$, $t = 0.67$, $P < 0.6$) (Fig. 5B). Higher 2-cell symmetry (corresponding to a low symmetry score) in embryos corresponds to a higher success rate ($r = -0.29 \pm 0.09$, $t = -3.22$, $P < 5 \times 10^{-3}$) (Fig. 5C). For developmental timing, the amount of time it takes for embryos to develop to 2-cell stage is negatively correlated with clinical success rates ($r = -0.54 \pm 0.08$, $t = -6.75$, $P < 5 \times 10^{-11}$) (Fig. 5D). Embryos with higher degree of fragmentation on Day 3 have lower success rates ($r = -0.31 \pm 0.08$, $t = -3.88$, $P < 5 \times 10^{-4}$) (Fig. 5E). The amount of time it takes for embryos to develop to blastocyst is negatively correlated with clinical success rates ($r = -0.27 \pm 0.14$, $t = -1.93$, $P < 0.06$) (Fig. 5F).

Amongst the features, 2-cell time, PN fade time, degree of fragmentation on Day 3, and cell symmetry are significantly correlated with the probability of embryos resulting in live births. ZP thickness during the 1-cell stage shows no significant correlations with clinical success rates. Time of blastulation shows no significant correlations, which may be due to the small sample

size ($n = 407$). Results from additional features are included in [Supplementary Fig. S6](#).

Discussion

In this study, we have built off of our prior works (Jang et al., 2023; Leahy et al., 2020; Lukyanenko et al., 2021) to develop and validate BlastAssist, a holistic pipeline to measure a comprehensive set of interpretable features in clinical IVF. Instead of using a black-box algorithm, BlastAssist outputs meaningful measurements of embryos that can be interpreted and corroborated by embryologists, which should be helpful in clinical decision making. These measurements are more detailed and quantitative than those currently used in standard clinical practice, and are therefore likely to assist IVF clinicians in selecting embryos for transfer and for freezing. BlastAssist is capable of measuring diverse features that are believed to be biologically and clinically relevant, including fertilization, cell symmetry, degree of fragmentation, developmental timing, and blastocyst expansion. Thus, BlastAssist provides a comprehensive and holistic means of evaluating embryos.

Automated measurements should be carefully validated before being deployed. We thus conducted and reported detailed comparisons between the pipeline outputs and measurements from a panel of human experts. The pipeline we developed has very high accuracies, performing comparably to, and in some cases outperforming human experts.

Embryo evaluations in actual clinical settings may differ from those performed in idealized situations. We thus further compared BlastAssist measurements to annotations previously made by embryologists during the course of routine IVF treatments. Since the clinical annotations only consist of single measurements for each embryo, we cannot quantify uncertainties for

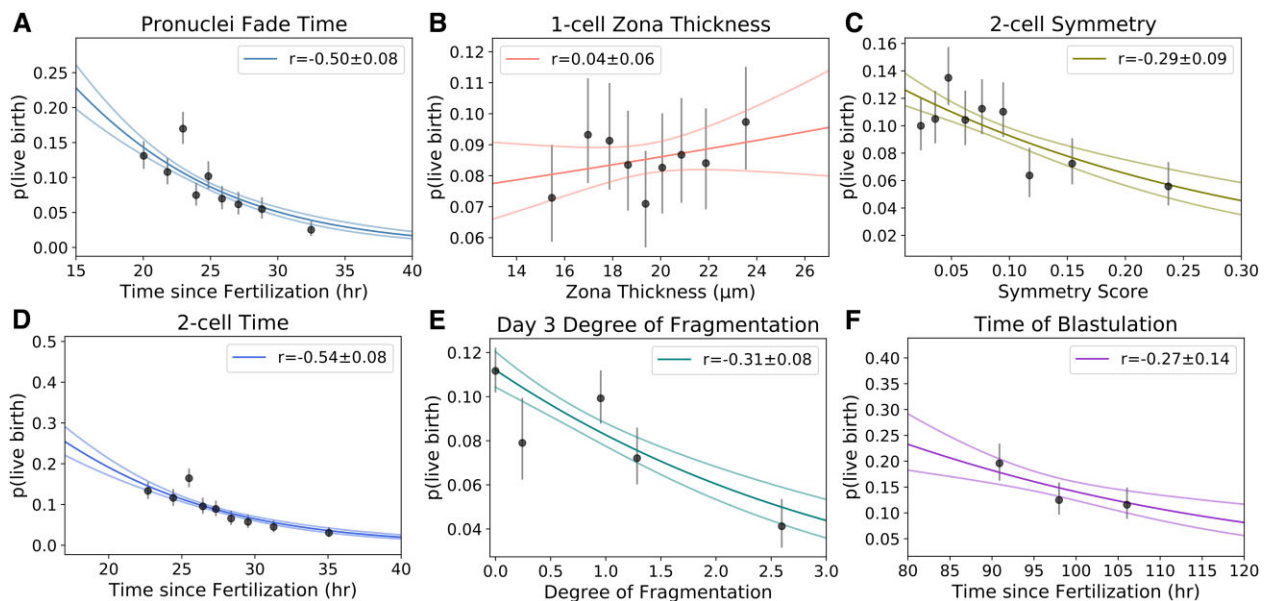


Figure 5. Associations of BlastAssist outputs to clinical live birth outcomes. Three thousand four hundred and twenty-one embryos (1801 cycles) were included in the association study. Estimated probability of an embryo resulting in a live birth as a function of (A) pronuclei (PN) fade time ($n = 3321$, $t = -6.25$, $P < 5 \times 10^{-10}$). (B) Zona pellucida (ZP) thickness during the 1-cell stage ($n = 3471$, $t = 0.67$, $P < 0.6$). (C) 2-cell symmetry score ($n = 3023$, $t = -3.22$, $P < 5 \times 10^{-3}$). (D) Time for embryos to reach 2-cell stage ($n = 3388$, $t = -6.75$, $P < 5 \times 10^{-11}$). (E) Degree of fragmentation on Day 3 ($n = 3251$, $t = -3.88$, $P < 5 \times 10^{-4}$). (F) Time for embryos to reach blastocyst stage ($n = 407$, $t = -1.93$, $P < 0.06$). The data points and error bars show the probability of the embryo producing live births estimated by a discrete model that fits an independent probability for each value. The curves and regions between faded lines show the best continuous nonlinear model fit with the data and its uncertainty. The correlations between $p(\text{live_birth})$ and the variables are plotted in each figure (see [Supplementary Data File S1: Correlations to live birth](#)).

these routine clinical measurements. However, the close agreement between BlastAssist and clinical annotations indicates that BlastAssist performs comparably to embryologist in a clinical setting.

To further investigate the potential of BlastAssist in a clinical setting, we studied the associations between BlastAssist measurements and clinical outcomes, including implantation results from SET cycles and live birth results from cycles with up to four embryos transferred. We found that PN fade time, but not ZP thickness, was significantly associated with the probability of embryos resulting in live births (Fig. 5A and B), while previous studies have given conflicting indications regarding the extent to which PN fade time (Lemmen et al., 2008; Coticchio et al., 2018; Barberet et al., 2019; Kobayashi et al., 2021) and ZP thickness (Gabielsen et al., 2000; Hagemann et al., 2010; Koifman et al., 2014; Lewis et al., 2017) are predictive of clinical outcomes. We found that cell symmetry was significantly positively correlated with the probability of embryos resulting in live births, while delayed cleavage time and degree of fragmentation were significantly negatively correlated with the probability of embryos resulting in live births (Fig. 5C–E), which is consistent with findings of previous studies regarding these variables (Ziebe et al., 2003; Della Ragione et al., 2007; Lemmen et al., 2008; Weitzman et al., 2010; Racowsky et al., 2011; Lee et al., 2012; Sela et al., 2012). Such associations should not form the sole basis of embryo selection algorithms since they do not account for extensive confounding factors (Rudin, 2019; Afnan et al., 2021a,b). However, incorporating BlastAssist outputs into mechanistic mathematical models, which explicitly account for patient factors and clinical practice (Leahy et al., 2021), has the potential to produce quantitative, predictive, and interpretable embryo selection algorithms.

Not only can BlastAssist be used for automated clinical evaluations to assist embryologists, it can also be used to obtain quantitative measurements to aid biomedical research. We used BlastAssist to analyze movies of 32 939 human embryos (67 043 973 images), resulting in an unprecedented dataset that will be a powerful resource for studying human pre-implantation embryology. Further studies of the data generated in this work may help to develop and test mathematical models of pre-implantation embryo development, which could further improve embryo selection. Ultimately, RCTs will be necessary to determine whether usage of the developed pipeline can improve success rates in clinical IVF.

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Data availability

The data that support the findings of this study are available from Lis Maternity Hospital Tel-Aviv Sourasky Medical Center but restrictions apply to the availability of these data due to regulations regarding protection of the rights and welfare of human subjects of research, and so are not publicly available. Data are however available from the authors (corresponding author: H.Y. Y.; Email: helen_yang@fas.harvard.edu) upon request and upon IRB approval and with Data Transfer Agreement with Lis Maternity Hospital Tel-Aviv Sourasky Medical Center, where the data were originally generated. The code used for this study has been deposited in a public GitHub repository (see <https://github.com/hyang185/BlastAssist>).

Acknowledgements

The authors would like to thank Yong Hyun Song and Rafael Elspas for their help with the association studies with live births. And additional thanks to all members of the Needleman, Ben-Yosef, and Hanspeter groups for their hard work and support.

Authors' roles

H.Y.Y. built the fragmentation scoring and blastocyst segmentation networks, performed integrated analysis of the dataset and correlation studies, performed human expert annotations, and wrote the manuscript. B.D.L. built the pronuclei detector and developmental stage classifier, performed the PN count and cell symmetry measurements, and performed human expert annotations. W.-D.J. built the blastomere detector and the pronuclei detector. D.W. discussed results. Y.K., R.R., A.C., and R.K. collected clinical images and annotations, and performed human expert annotations. F.A. supervised the clinical aspect of the project. M. V., C.P.K., and L.C. performed human expert annotations. H.P. supervised the computer vision aspect of the project. D.J.N. and D. B.-Y. edited the manuscript and supervised the entire project.

Funding

This work was supported by Harvard Quantitative Biology Initiative, the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (award number 1764269), the National Institute of Health (award number R01HD104969), and the Sagol fund for embryos and stem cells as part of the Sagol Network.

Conflict of interest

The authors declare no conflicts of interest.

References

- Adamson G, Zegers-Hochschild F, Dyer S, Chambers G, de Mouzon J, Ishihara O, Kupka M, Banker M, Jwa S, Elgindy E et al. International committee for monitoring assisted reproductive technology: world report on assisted reproductive technology, 2018. 2022. <https://www.icmartivf.org/wp-content/uploads/ICMART-ESHRE-WR2018-Preliminary-Report.pdf> (13 February 2024, 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* 2021a;**4**:hoab040.
- Afnan MAM, Rudin C, Conitzer V, Savulescu J, Mishra A, Liu Y, Afnan M. Ethical implementation of artificial intelligence to select embryos in in vitro fertilization. In: 2021 AAAI/ACM Conference on AI, Ethics, and Society. Virtual. 2021b, 316–326. <https://doi.org/10.1145/3461702.3462589>.
- Ahlström A, Lundin K, Lind AK, 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.
- Amir H, Barbash-Hazan S, Kalma Y, Frumkin T, Malcov M, Samara N, Hasson J, Reches A, Azem F, Ben-Yosef D. Time-lapse imaging reveals delayed development of embryos carrying unbalanced chromosomal translocations. *J Assist Reprod Genet* 2019; **36**:315–324.

- 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.
- Barberet J, Bruno C, Valot E, Antunes-Nunes C, Jonval L, Chammas J, Choux C, Ginod P, Sagot P, Soudry-Faure A et al. Can novel early non-invasive biomarkers of embryo quality be identified with time-lapse imaging to predict live birth? *Hum Reprod* 2019; **34**:1439–1449.
- Benesty J, Chen J, Huang Y. On the importance of the pearson correlation coefficient in noise reduction. *IEEE Trans Audio Speech Lang Process* 2008; **16**:757–765.
- 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.
- Breu H, Gil J, Kirkpatrick D, Werman M. Linear time Euclidean distance transform algorithms. *IEEE Trans Pattern Anal Machine Intell* 1995; **17**:529–533.
- Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Hickman CFL. Modelling a risk classification of aneuploidy in human embryos using non-invasive morphokinetics. *Reprod Biomed Online* 2013; **26**:477–485.
- Campbell A, Fishel S, Laegdsmand M. Aneuploidy is a key causal factor of delays in blastulation: author response to ‘a cautionary note against aneuploidy risk assessment using time-lapse imaging’. *Reprod Biomed Online* 2014; **28**:279–283.
- Centers for Disease Control and Prevention (CDC). 2019 assisted reproductive technology fertility clinic and national summary report. US Dept of Health and Human Services, 2021. https://archive.cdc.gov/www_cdc_gov/art/reports/2019/pdf/2019-Report-ART-Fertility-Clinic-National-Summary-h.pdf (13 February 2024, date last accessed).
- Chavez-Badiola A, Farias AFS, 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* 2020; **10**:4394–4396.
- Coticchio G, Mignini Renzini M, Novara P, Lain M, De Ponti E, Turchi D, Fadini R, Dal Canto M. Focused time-lapse analysis reveals novel aspects of human fertilization and suggests new parameters of embryo viability. *Hum Reprod* 2018; **33**:23–31.
- De Neubourg D, Gerris J. Single embryo transfer – state of the art. *Reprod Biomed Online* 2003; **7**:615–622.
- Della Ragione T, Verheyen G, Papanikolaou EG, Van Landuyt L, Devroey P, Van Steirteghem A. Developmental stage on day-5 and fragmentation rate on day-3 can influence the implantation potential of top-quality blastocysts in IVF cycles with single embryo transfer. *Reprod Biol Endocrinol* 2007; **5**:2–8.
- Dolinko AV, Farland L, Kaser D, Missmer S, Racowsky C. National survey on use of time-lapse imaging systems in IVF laboratories. *J Assist Reprod Genet* 2017; **34**:1167–1172.
- Ebner T, Moser M, Sommergruber M, Tews G. Selection based on morphological assessment of oocytes and embryos at different stages of preimplantation development: a review. *Hum Reprod Update* 2003; **9**:251–262.
- Freedman D, Pisani R, Purves R. *Statistics (International Student Edition)*, 4th edn. New York: WW Norton & Company, 2007.
- Gabrielsen A, Bhatnager PR, Petersen K, Lindenberg S. Influence of zona pellucida thickness of human embryos on clinical pregnancy outcome following in vitro fertilization treatment. *J Assist Reprod Genet* 2000; **17**:323–328.
- Hagemann AR, Lanzendorf SE, Jungheim ES, Chang AS, Ratts VS, Odem RR. A prospective, randomized, double-blinded study of assisted hatching in women younger than 38 years undergoing in vitro fertilization. *Fertil Steril* 2010; **93**:586–591.
- Harun MY, Huang T, Ohta AT. Inner cell mass and trophectoderm segmentation in human blastocyst images using deep neural network. In: 2019 IEEE 13th International Conference on Nano/Molecular Medicine & Engineering (NANOMED), Gwangju, Korea. New York: IEEE, 2019, 214–219.
- Hoffman R, Gross L. Modulation contrast microscope. *Appl Opt* 1975; **14**:1169–1176.
- Jang WD, Wei D, Zhang X, Leahy B, Yang H, Tompkin J, Ben-Yosef D, Needleman D, Pfister H. Learning vector quantized shape code for amodal blastomere instance segmentation. In: IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia. New York: IEEE, 2023, 1–5.
- Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015; **349**:255–260.
- Khan A, Gould S, Salzmann M. Deep convolutional neural networks for human embryonic cell counting. In: European Conference on Computer Vision, Amsterdam, The Netherlands. Cham: Springer, 2016; Part I (14):339–348.
- Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, Sigaras A, Lavery S, Cooper LA, Hickman C et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ Digit Med* 2019; **2**:21–29.
- Kirkegaard K, Agerholm IE, Ingerslev HJ. Time-lapse monitoring as a tool for clinical embryo assessment. *Hum Reprod* 2012; **27**:1277–1285.
- Kobayashi T, Ishikawa H, Ishii K, Sato A, Nakamura N, Saito Y, Hasegawa H, Fujita M, Mitsushashi A, Shozu M. Time-lapse monitoring of fertilized human oocytes focused on the incidence of Opn embryos in conventional in vitro fertilization cycles. *Sci Rep* 2021; **11**:18862.
- Koifman M, Lahav-Baratz S, Shopen L, Idit B, Ishai D, Wiener-Megnazi Z, Auslender R, Dirnfeld M. In vitro fertilization outcomes following assisted hatching of embryos with thick zona pellucida—a prospective randomized study. *Adv Reprod Sci* 2014; **02**:76–82.
- Kragh MF, Rimestad J, Berntsen J, Karstoft H. Automatic grading of human blastocysts from time-lapse imaging. *Comput Biol Med* 2019; **115**:103494.
- Lau T, Ng N, Gingold J, Desai N, McAuley J, Lipton ZC. Embryo staging with weakly-supervised region selection and dynamically-decoded predictions. In: Machine Learning for Healthcare Conference, Ann Arbor, MI, USA. Cambridge, MA, USA: PMLR, 2019, 663–679.
- Leahy BD, Jang WD, Yang HY, Struyven R, Wei D, Sun Z, Lee KR, Royston C, Cam L, Kalma Y et al. Automated measurements of key morphological features of human embryos for IVF. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru. Cham: Springer, 2020; Part V(23):25–35.
- Leahy BD, Racowsky C, Needleman D. Inferring simple but precise quantitative models of human oocyte and early embryo development. *J R Soc Interface* 2021; **18**:20210475.
- Lee AM, Connell MT, Csokmay JM, Styer AK. Elective single embryo transfer—the power of one. *Contracept Reprod Med* 2016; **1**:1–7.
- Lee E, Illingworth P, Wilton L, Chambers GM. The clinical effectiveness of preimplantation genetic diagnosis for aneuploidy in all 24 chromosomes (PGD-A): systematic review. *Hum Reprod* 2015; **30**:473–483.
- Lee MJ, Lee RKK, Lin MH, Hwu YM. Cleavage speed and implantation potential of early-cleavage embryos in IVF or ICSI cycles. *J Assist Reprod Genet* 2012; **29**:745–750.

- Lemmen J, Agerholm I, Ziebe S. Kinetic markers of human embryo quality using time-lapse recordings of IVF/ICSI-fertilized oocytes. *Reprod Biomed Online* 2008;**17**:385–391.
- Lewis EI, Farhadifar R, Farland LV, Needleman D, Missmer SA, Racowsky C. Use of imaging software for assessment of the associations among zona pellucida thickness variation, assisted hatching, and implantation of day 3 embryos. *J Assist Reprod Genet* 2017;**34**:1261–1269.
- Lukyanenko S, Jang WD, Wei D, Struyven R, Kim Y, Leahy B, Yang H, Rush A, Ben-Yosef D, Needleman D et al. Developmental stage classification of embryos using two-stream neural network with linear-chain conditional random field. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2021: 24th International Conference*, Strasbourg, France. Cham: Springer, 2021; Part VIII(24):363–372.
- Malmsten J, Zaninovic N, Zhan Q, Rosenwaks Z, Shan J. Automated cell stage predictions in early mouse and human embryos using convolutional neural networks. In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, Chicago, IL, USA. New York: IEEE, 2019, 1–4.
- Mastenbroek S, Scriven P, Twisk M, Viville S, Van der Veen F, Repping S. What next for preimplantation genetic screening? More randomized controlled trials needed? *Hum Reprod* 2008;**23**:2626–2628.
- Mastenbroek S, van der Veen F, Aflatoonian A, Shapiro B, Bossuyt P, Repping S. Embryo selection in IVF. *Hum Reprod* 2011;**26**:964–966.
- McHugh ML. The chi-square test of independence. *Biochem Med (Zagreb)* 2013;**23**:143–149.
- Norwitz ER, Edusa V, Park JS. Maternal physiology and complications of multiple pregnancy. *Semin Perinatol* 2005;**29**:338–348.
- Paternot G, Wetsels AM, Thonon F, Vansteenbrugge A, Willemen D, Devroe J, Debrock S, D'Hooghe TM, Spiessens C. Intra- and inter-observer analysis in the morphological assessment of early stage embryos during an IVF procedure: a multicentre study. *Reprod Biol Endocrinol* 2011;**9**:1–5.
- Paulson RJ. Hidden in plain sight: the overstated benefits and underestimated losses of potential implantations associated with advertised PGT-A success rates. *Hum Reprod* 2020;**35**:490–493.
- Racowsky C, Stern JE, Gibbons WE, Behr B, Pomeroy KO, Biggers JD. National collection of embryo morphology data into society for assisted reproductive technology clinic outcomes reporting system: associations among day 3 cell number, fragmentation and blastomere asymmetry, and live birth rate. *Fertil Steril* 2011;**95**:1985–1989.
- Rad RM, Saeedi P, Au J, Havelock J. Multi-resolutional ensemble of stacked dilated u-net for inner cell mass segmentation in human embryonic images. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece. New York: IEEE, 2018, 3518–3522.
- Rad RM, Saeedi P, Au J, Havelock J. Trophoctoderm segmentation in human embryo images via inceptioned U-Net. *Med Image Anal* 2020;**62**:101612.
- Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;**1**:206–215.
- Sela R, Samuelov L, Almog B, Schwartz T, Cohen T, Amit A, Azem F, Ben-Yosef D. An embryo cleavage pattern based on the relative blastomere size as a function of cell number for predicting implantation outcome. *Fertil Steril* 2012;**98**:650–656.e4.
- Sfakianoudis K, Maziotis E, Grigoriadis S, Pantou A, Kokkini G, Trypidi A, Giannelou P, Zikopoulos A, Angeli I, Vaxevanoglou T et al. Reporting on the value of artificial intelligence in predicting the optimal embryo for transfer: a systematic review including data synthesis. *Biomedicine* 2022;**10**:697.
- Silver DH, Feder M, Gold-Zamir Y, Polsky AL, Rosentraub S, Shachor E, Weinberger A, Mazur P, Zukin VD, Bronstein AM. *Data-driven prediction of embryo implantation probability using ivf time-lapse imaging*. Montréal, QC, Canada: Medical Imaging with Deep Learning, 2020. <https://openreview.net/forum?id=TujK1uTkTP> (13 February 2024, date last accessed).
- Simopoulou M, Sfakianoudis K, Maziotis E, Antoniou N, Rapani A, Anifandis G, Bakas P, Bolaris S, Pantou A, Pantos K et al. Are computational applications the “crystal ball” in the IVF laboratory? The evolution from mathematics to artificial intelligence. *J Assist Reprod Genet* 2018;**35**:1545–1557.
- Student. The probable error of a mean. *Biometrika* 1908;**6**:1–25.
- 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.
- van Montfoort AP, Dumoulin JC, Land JA, Coonen E, Derhaag JG, Evers JL. Elective single embryo transfer (eSET) policy in the first three IVF/ICSI treatment cycles. *Hum Reprod* 2005;**20**:433–436.
- VerMilyea M, Hall J, Diakiw S, Johnston A, Nguyen T, Perugini D, Miller A, Picou A, Murphy A, 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.
- Weitzman VN, Schnee-Riesz J, Benadiva C, Nulsen J, Siano L, Maier D. Predictive value of embryo grading for embryos with known outcomes. *Fertil Steril* 2010;**93**:658–662.
- Ziebe S, Lundin K, Loft A, Bergh C, Nyboe Andersen A, Selleskog U, Nielsen D, Grøndahl C, Kim H, Arce JC; CEMAS II and Study Group. FISH analysis for chromosomes 13, 16, 18, 21, 22, X and Y in all blastomeres of IVF pre-embryos from 144 randomly selected donated human oocytes and impact on pre-embryo morphology. *Hum Reprod* 2003;**18**:2575–2581.