Enhancing Clinical IVF Embryo Selection through the Integration of Artificial Intelligence and Bayesian Statistics

HARVARD
Kenneth C. Griffin

GRADUATE SCHOOL
OF ARTS AND SCIENCES

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Committee on Higher Degrees in Biophysics

have examined a dissertation entitled

Enhancing Clinical IVF Embryo Selection through the Integration of Artificial Intelligence and Bayesian Statistics

presented by **Yu Helen Yang**

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Rachelle Gaudet

Signature _____

Typed name: Prof. Hanspeter Pfister

Signature _____

Typed name: Prof. Catherine Racowsky

Date: ____April 8, 2024_____

Enhancing Clinical IVF Embryo Selection through the Integration of Artificial Intelligence and Bayesian Statistics

A DISSERTATION PRESENTED
BY
YU HELEN YANG
TO
THE DEPARTMENT OF BIOPHYSICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOPHYSICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 8 2024

Thesis advisor: Professor Daniel J. Needleman Yu Helen Yang

Enhancing Clinical IVF Embryo Selection through the Integration of Artificial Intelligence and Bayesian Statistics

Abstract

Despite major advancements in IVF technologies, the success rate of IVF remains low at around 35% and the main challenge of clinical IVF is embryo selection. Our approach to improve embryo selection leverages the advantages of AI, its ability to analyzing large and complex datasets, and Bayesian statistics, its ability to robustly infer causal relationships in complex problems. First, we enhanced the precision and efficiency of Time-Lapse Microscopy (TLM) image analysis through the implementation of Computer Vision (CV) techniques. We developed AI algorithms to automate the extraction of morphokinetic features from TLM movies of human embryos, providing a more objective evaluation of embryo quality. Second, we integrated these CV networks into BlastAssist, a pipeline designed to measure comprehensive, quantitative, and clinically-relevant features in IVF. To validate our pipeline, we conducted detailed comparisons between BlastAssist measurements and manual assessments by human experts, annotations from embryologists during routine treatments, outcomes of single embryo transfer (SET) cycles, and live birth outcomes of transferred embryos. Lastly, using the unprecedentedly large dataset generated by the BlastAssist pipeline in conjunction with electronic health record (EHR) data from three IVF labs, we constructed probabilistic graphical models (PGMs) for the complex IVF process across three key stages: ovarian stimulation, fertilization, and embryo development. These graphical models elucidate causal relationships among variables in clinical IVF, providing valuable insights for directing future clinical research. Overall, this project has the potential for significant clinical impact. The BlastAssist pipeline has the potential as a powerful tool to streamline the IVF image analysis process and assist embryologists in embryo selection. The integration of AI and Bayesian statistics can enhance our understanding of this complex process and help us identify key clinical predictors and improve clinical outcomes.

Contents

List of Authors

The following authors contributed to Chapter 1:

B. D. Leahy, W. D. Jang, Y. H. Yang, R. Struyven, D. Wei, Z. Sun, K. R. Lee, C. Royston, L. Cam, Y. Kalma, F. Azem, D. Ben-Yosef, H. Pster, and D. J. Needleman.

The following authors contributed to Chapter 2:

Y. H. Yang, B. D. Leahy, W. D. Jang, D. Wei, Y. Kalma, R. Rahav, A. Carmon, R. Kopel, F. Azem, M. Venturas, C. P. Kelleher, L. Cam, H. Pster, D. J. Needleman, and D. Ben-Yosef.

The following authors contributed to Chapter 3: Y. H. Yang, B. D. Leahy, Y. Kalma, D. Ben-Yosef, and D. J. Needleman.

Listing of figures

List of Tables

Acknowledgments

I EXTEND MY DEEPEST GRATITUDE TO:

Professor Daniel J. Needleman for his exceptional guidance and support as my advisor.

Professors Dalit Ben-Yosef, Hanspeter Pfister and Catherine Racowsky for their collaboration and invaluable feedback.

Professors Olivier Pourquié and Rachelle Gaudet for their insightful contributions as committee members.

Colleagues Brian D. Leahy and Won-Dong Jang for their significant contributions.

Michele R. Jakoulov and the Biophysics Program for their support.

Everyone involved in the project within the groups of Professors Needleman, Ben-Yosef, and Pfister.

Members of the Needleman lab for their enduring support.

Thank you all for your invaluable contributions to this work.

0

Introduction

0.1 Background

Overview of *in vitro* fertilization (IVF) and its significance

Infertility is a widespread problem that affects a significant portion of the population worldwide[1]. According to the Centers for Disease Control and Prevention (CDC), approximately 1 in 6 couples experience infertility at some point of their reproductive life, and 1 in 3 couples if maternal age is older than 35 years old[22,21], making it a common issue that impacts millions of people worldwide. Of all the assisted reproductive technologies (ART) available, *in vitro* fertilization (IVF) is the most effective treatment for infertility[21]. In clinical IVF, patients first undergo ovarian stimulation to produce multiple oocytes, and these oocytes are retrieved by embryologists and fertilized *in vitro*. The successfully fertilized embryos are then cultured *in vitro* for up to 5 days. Embryologists then select the best embryo and transfer it to the uterus, where it can implant and develop into a healthy fetus.

The success rate of IVF varies depending on many factors, including the age of the woman, the quality of the sperm, and the skill of the embryologist, etc[119,23,25]. Global reports of clinical outcomes show that the success rate of IVF still remains low, at around 35%, with a higher success rate for women under 35 years old[21,1]. Despite advances in many aspects of IVF techniques, embryo selection remain a key challenge[60,55]. Selecting the most viable embryos is a highly complicated problem due to the complex clinical environments with many confounding factors. In addition, there is still a lack of understanding of human pre-implantation embryo development, as it is extremely difficult to study human embryos in an experimental setting due to ethics reasons[8,56].

Embryologists tend to transfer more embryos for patients with poor prognosis or low embryo qualities, hoping to improve pregnancy rates. However, increasing the number of embryos transferred also increases the risks associated with multiple pregnancies, such as premature birth and maternal complications[81]. To address these concerns, there is a growing emphasis on single embryo

transfer (SET), where only the best embryo is selected and transferred to maximize the chances of a successful pregnancy while minimizing risks of multiple pregnancies[65,29,13]. Making this major shift towards SET will require continued research and advancement in the field of IVF, especially significant improvement on embryo selection techniques, as well as a deeper understanding of the complex factors that influence pre-implantation embryo development.

Challenges in current clinical IVF embryo selection methods

The current standard of care in embryo selection is based on morphological grading[77,37], where embryologists assess the morphology of the oocytes or embryos manually under a microscope. The goal of morphological grading is to identify any qualitative abnormalities or irregularities in the shape, structure or development of the embryos, which are associated to their potential for implantation and pregnancy. Despite the improvements in morphological grading, there are still several challenges associated with it, including: 1). Inter and intra -observer variability: Morphological grading is subjective, and there can be variations in the interpretation of abnormalities between different embryologists and with the same embryologists[83]. 2). Limited predictive value: Morphological grading is not always accurate in predicting the outcome of an IVF cycle[117]. Many abnormalities and features evaluated may not actually affect the implantation potential of the eggs or embryos[112] 3). Time-consuming: Manual morphological grading can be time-consuming, which can increase the burden of embryologists. 4). Limited assessment: Manual morphological grading only assess the embryos at a few time points, and may not provide a complete picture of their quality.

Pre-implantation Genetic Testing for Aneuploidies (PGT-A) is becoming more common in the field of IVF as a means of embryo selection, where a few cells are biopsied and tested for aneuploidy[49]. While there is strong correlation between aneuploidy and lower viability[122], it's important to recognize that this test is not without controversy. Here are some of the challenges and controversies surrounding PGT-A: 1). Invasiveness: PGT-A involves the biopsy of a small number of cells

from the embryo, which can be invasive and carries a risk of harm to the embryo[97]. 2). Efficacy: The efficacy of PGT-A in detecting aneuploidy is not 100%, and some embryos that test aneuploid may still implant and result in a healthy pregnancy, likely due to the prevalence of mosaicism in pre-implantation human embryos[87,115]. 3). Cost: PGT-A is an expensive test, and not widely available in clinics, which can be a barrier for some patients[104]. 4). Impact on pregnancy outcomes: Some studies have suggested that PGT-A may not improve pregnancy outcomes, and that other factors such as maternal age and embryo quality may play a more important role[97,66,84,76].

Time-lapse microscopy (TLM) is a technique that uses a specialized microscope to capture high-resolution images of embryos at regular time intervals, typically every 10-30 minutes, over a period of several days. This allows researchers to observe the dynamic processes of embryonic development in real-time, providing valuable insights into the growth and development of embryos. One of the key advantages of TLM is that it provides a much more detailed and accurate assessment of embryo development compared to traditional manual morphological grading. TLM has been adopted in many IVF clinics for the culture of embryos[35,9,55]. TLM systems has shown promise in improving the accuracy and efficiency of embryo grading, but there are still limitations to their effectiveness. Many clinics that utilizes these systems still rely on manual evaluation of embryo movies, which can be time-consuming and subjective. A recent randomized controlled trial (RCT) found that the current method of utilizing TLM did not lead to improved IVF outcomes compared to manual morphological grading alone[4]. This may be due to the fact that not all of the information from TLM is being leveraged, and there is still a need for further development and refinement of these systems to improve their accuracy and applicability in clinical settings.

With the recent rapid advancements in artificial intelligence (AI), which are based on constructing predictive models from the (often subtle) associations present in 'training' data, researchers have attempted to develop AI algorithms for use in embryo evaluation and selection. Some previous AI algorithms were developed to directly predict clinical outcomes from clinical images or

movies[16,102,24,114,108]. These algorithms are referred to as "black-box" algorithms, as their inner workings are not transparent or interpretable. In other words, the algorithm's decision-making process is opaque, and it is difficult to understand why it makes certain predictions or decisions. These "black-box" algorithms are often used in applications where interpretability is not a primary concern, such as image or speech recognition, natural language processing, and other complex predictive tasks. However, in the IVF and other medical settings, interpretability is crucial in clinical decision making[36,86]. Therefore, an alternative means of AI algorithm with high interpretability is necessary to for it to be applicable in clinical IVF[2,96,3].

0.2 Aims

Aim 1. Improve the accuracy and efficiency of TLM image analysis using Computer Vision

Our objective is to build a clinical selection algorithm to fully leverage the advantages of both TLM systems and AI, without compromising on interpretability of the model. We first aim to improve the accuracy and efficiency of image analysis of TLM movies using computer vision (CV) methods. CV is a field of AI that deals with enabling computers to interpret and understand visual information from the world around us. This includes images, videos, and other types of visual data. The goal of CV is to enable AI algorithms to perform tasks that would typically require human visual perception, such as segmentation, object detection and recognition, object understanding (classification).

Traditional methods of IVF image analysis rely on manual observation and measurement of embryos, which can be time-consuming and subjective. However, by leveraging the power of CV, we aim to automatically extract comprehensive morphokinetic features from movies of embryos during the early stages of development. These features can provide a more accurate and objective assessment of embryo quality, enabling embryologists to make more informed decisions about which

embryos to select for transfer. We hope that this will ultimately lead to improved clinical success rates and enhanced patient experience. Our approach will also streamline the image analysis process of TLM, reducing the time and cost associated with manual analysis, and enabling clinics to provide better care to more patients. By combining cutting-edge CV techniques with the expertise of embryologists, we believe that we can revolutionize the field of IVF analysis and improve the chances of success for couples undergoing fertility treatment.

Aim 2. Measure Interpretable and clinically-relevant features of Human Embryos

Before we can deploy the AI algorithm we developed in Aim 1 in clinical settings, it is crucial that we measure key clinical features and conduct detailed comparisons between our outcomes and the measurements of human experts and clinicians.

In order to corroborate the algorithms we developed, we first combine the CV networks we developed into BlastAssist, a pipeline that can measure interpretable and clinically-relevant features of human embryos. 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 then compared the pipeline measurements of individual images/embryos to manual measurements by human experts for sets of features, including: 1) fertilization status, 2) cell symmetry, 3) degree of fragmentation, 4) developmental timing. We then conducted detailed comparisons between pipeline outputs and annotations made by embryologists during routine treatments for features, including: 1) fertilization status, 2) pronuclei (PN) fade time, 3) degree of fragmentation on day 2, 4) time of blastulation. In addition, we compared the pipeline outputs to the outcomes of single embryo transfer (SET) cycles with known implantation results, and to the live birth results of embryos that were transferred.

The ultimate goal of BlastAssist is to provide embryologists with a powerful tool that can aid in the embryo selection process, ultimately improving the clinical success rate.

Aim 3. Build Causal Models of Clinical IVF Features

The clinical IVF environment is extremely complicated, with many variables at play, from patient data, treatment data, to embryo data, and IVF outcomes, and many other confounding variables. Most clinical IVF studies have used statistics to find correlations between variables, hoping to gain understanding of this complex process. However, there are two major problems: 1) Amongst the large number variables in clinical IVF, almost all variables are inter-correlated with each other, making understanding and modeling the process difficult. And the frequentist statistics that most studies use is not suitable to study these complex systems. This is particularly important in IVF, where many confounding variables can significantly affect pregnancy outcomes. 2) most fundamentally, typical correlation studies are aimed 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?

Overall, Bayesian causal models offer a powerful tool for clinical decision-making in IVF, allowing for more accurate predictions of treatment outcomes while incorporating prior knowledge and accounting for confounding variables.

0.3 Significance

This project has the potential for significant clinical impact. First, the BlastAssist pipeline can not only save valuable time for the embryologists, but also increase consistency and accuracy for clinical annotations. The pipeline can be easily incorporated into clinical settings and serve as a tool to assist embryologists in clinical decision making and improve clinical efficiency.

Second, we are able to obtain an unprecedented amount of human pre-implantation developmental data from the pipeline measurements. Using this data, one can potentially answer many

significant biological questions, and conduct retrospective clinical trials. We are also able to study and potentially answer other fundamental questions of biology and physics during the human pre-implantation embryonic development but these will be beyond the scope of my projects.

In addition, the probabilistic graphical models we built using Bayesian statistics will hopefully allow us to improve our current embryo selection process in clinical IVF. Instead of using AI to produce ”black-box” algorithms to directly predict implantation outcome from images and videos, we leverage all the information from the TLM videos and the advantages of AI, and built causal inference models.

The integration of AI and Bayesian statistics has the potential to significantly enhance the accuracy of embryo selection in IVF. By integrating CV algorithms into the BlastAssist pipeline to analyze large datasets, embryologist can identify subtle patterns and trends that may not be apparent through traditional methods. Additionally, Bayesian statistics can help to incorporate prior knowledge and uncertainty into the selection process, leading to more reliable predictions and better outcomes.

1

Automating IVF Analysis: Morphokinetic Assessment of Human Embryos through Computer Vision

This chapter is adapted from the manuscript of Automated Measurements of Key Morphological Features of Human Embryos for IVF[63].

1.1 Introduction

One in six couples worldwide suffer from infertility[28]. Many of those couples seek to conceive via In-Vitro Fertilization (IVF). In IVF, a patient is stimulated to produce multiple oocytes. The oocytes are retrieved, fertilized, and the resulting embryos are cultured *in vitro*. Some of the embryos are then transferred to the mother's uterus in the hopes of achieving a pregnancy; the remaining viable embryos are cryopreserved for future treatments.

While transferring multiple embryos to the mother increases the potential for success, it also increases the potential for multiple pregnancies, which are strongly associated with increased risks of maternal morbidity and offspring morbidity and mortality[81]. Thus, it is highly desirable to transfer only one embryo, to produce only one healthy child[82]. This requires clinicians to select the highest quality embryos for transfer, which remains highly challenging[88].

The current standard of care is to select embryos primarily based on their morphology, by examining them under a microscope. In a typical embryo, after fertilization the two pronuclei, which contain the father's and mother's DNA, move together and migrate to the center of the embryo. The embryo undergoes a series of cell divisions, during the "cleavage stage." Four days after fertilization, the embryo compacts and the cells firmly adhere to each other, at which time it is referred to as a compact "morula." On the fifth day, the embryo forms a "blastocyst", consisting of an outer layer of cells (the trophectoderm) enclosing a smaller mass (the inner-cell mass). On the sixth day, the blastocyst expands and hatches out of the zona pellucida (the thin eggshell that surrounds the embryo)[38]. Clinicians score embryos by manually measuring features such as cell number, cell shape,

cell symmetry, the presence of cell fragments, and blastocyst appearance[38], usually at discrete time points. Recently, many clinics have started to use time-lapse microscopy systems that continuously record movies of embryos without disturbing their culture conditions[95,35,9]. However, these videos are typically analyzed manually, which is time-consuming and subjective.

Previous researchers have trained convolutional neural networks (CNNs) to directly predict embryo quality, using either single images or time-lapse videos[85,108]. However, interpretability is vital for clinicians to make informed decisions on embryo selection, and an algorithm that directly predicts embryo quality from images is not interpretable. Worse, since external factors such as the patient age[40] and body-mass index[18] also affect the success of an embryo transfer, an algorithm trained to predict embryo quality may end up learning a representation of confounding variables that may change as IVF practices or demographics change. Some researchers have instead trained CNNs to extract a few identifiable features from images of embryos, such as blastocyst size[53], blastocyst grade[61,39,54], cell boundaries[89], or the number of cells when there are 4 or fewer[52,62]. While obviating any problems with interpretability, these works leave out many key features that are believed to be important for embryo quality.

Here, we present a machine learning suite of five CNNs to extract key morphokinetic features from time-lapse videos of embryos in clinical IVF. We work closely with clinicians to measure biological features relevant for clinical IVF: segmentation of the zona pellucida (Fig. 1.1a), grading the degree of fragmentation (Fig. 1.1b), classification of the developmental stage from 1-cell to blastocyst (Fig. 1.1c), object instance segmentation of cells during cleavage stage (Fig. 1.1d), and object instance segmentation of pronuclei before the first cell division (Fig. 1.1e). With the exception of segmenting the zona pellucida, all of these features have been used for embryo selection[5,88,7,80]; we segment the zona pellucida both to improve the other classifiers and because zona properties occasionally inform other IVF procedures[26]. The five CNNs work together in a unified pipeline, combining results from one another to improve performance.

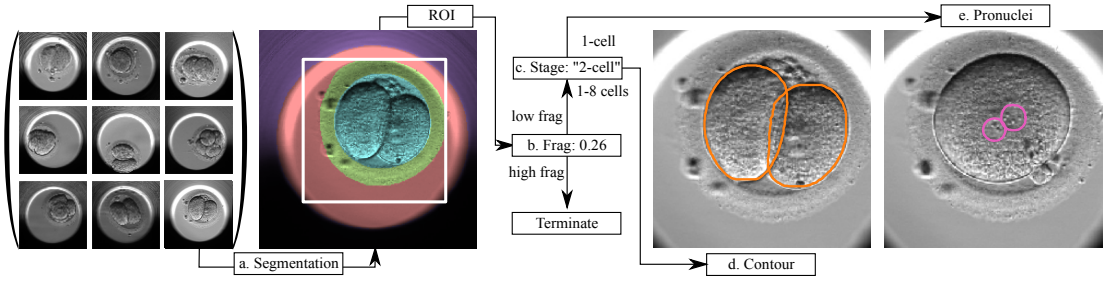

Figure 1.1: Computer vision pipeline for extracting morphokinetic features from human embryos. We first segment the image to locate the embryo (panel a, regions colored according to segmentation). The segmentation provides a region-of-interest (ROI, white box) for the other 4 classifiers, starting with embryo fragmentation (b); the image shown has a predicted fragmentation score of 0.26. If the embryo's fragmentation score is less than 1.5, we classify the developmental stage (c); this image is classified as a 2-cell embryo. We then detect cells in cleavage stage embryos (orange contours in d) and pronuclei in 1-cell embryos (magenta contours in e).

1.2 Dataset

We train the CNNs using data from the , the most widely-used system for IVF with standardized, time-lapse microscopy. images are grayscale, and taken using Hoffman Modulation Contrast microscopy[48], in which the intensity roughly corresponds to refractive index gradients. In our dataset, the takes an image every 20 minutes at 7 focal planes, usually at 15 increments. The embryos are recorded for 3 – 5 days, corresponding to 200 – 350 images at each focal plane (*i.e.* 1400 – 2450 images per embryo), although embryos are occasionally removed from the incubation system for clinical procedures. To train the CNNs, we curate a dataset with detailed, frame-by-frame labels for each task.

1.3 Our Pipeline

For each time-lapse video, we measure 5 morphokinetic features using 5 networks:

Zona Pellucida Segmentation We first perform semantic segmentation to identify regions of the embryo, segmenting the image into four regions: pixels outside the well, inside the well, the

zona pellucida, and the space inside the zona pellucida (the perivitelline space and embryo; Figure 1.2, left). We segment the images by using a fully-convolutional network (FCN[71]; based on Resnet101 [47]) to predict a per-pixel class probability for each pixel in the image. We train the FCN with images chosen from 203 embryos at 3,618 time points.

The first network in our pipeline, zona segmentation, takes the full 500×500 pixel image as input. We use the zona segmentation to crop the images to 328×328, centered around the embryo, as input for the other networks.

Fragmentation Scoring With the cropped image from the zona pellucida segmentation, we score the embryo's degree of fragmentation, using a regression CNN (InceptionV3 [106]). For each time point in the movie we analyze, we run the CNN on the three middle focal planes (Figure 1.3, left). The network takes each single-focus image as input and predicts a fragmentation score of 0 (0% fragments), 1 ($<$10%), 2 (10-20%), or 3 ($\geq$20%), following clinical practice. We train the network to minimize the L^1 loss on cleavage-stage images of 989 embryos at 16,315 times, each labeled with an integer score from $0-3$; we use a validation set of 205 embryos labeled at 3,416 times for early stopping[43]. We produce a final score by averaging the three separate scores from the three separate focal planes.

Counting and identifying cells in fragmented embryos is difficult, inhibiting the labeling of train or test data for these embryos. Moreover, since high fragmentation is strongly correlated with low embryo viability[5], in standard clinical practice highly fragmented embryos are frequently discarded. Thus, we only train the rest of the networks on embryos with less than 15% fragmentation.

Stage Classification For low fragmentation embryos, we classify the embryo's developmental stage over time using a classification CNN (ResNeXt101 [118]). The classifier takes the three middle focal planes as input and predicts a 13-element vector of class probabilities, with 9 classes for cleavage-stage embryos (one each for 1–8 cells and one for $\geq$ 9 cells) and one class each for morula (M), blastocyst (B), empty wells (E), and degenerate embryos (D; Figure 1.4, left). To account for

inaccuracies in the training data labels, we trained the classifier with a soft loss function modified from the standard cross-entropy loss

$$\log\left(p(\ell|m)\right) = \log\left(\sum_t p(\ell|t)p(t|m)\right) \quad , \tag{1.1}$$

where t is the true stage of an image, ℓ the (possibly incorrect) label, and m the model's prediction. We measured $p(\ell|t)$ by labeling 23,950 images in triplicate and using a majority vote to estimate the true label t of each image. This soft-loss differs from the regularized loss in [106] by differentially weighting classes; for instance, $p(\ell = \text{1-cell}|t = \text{1-cell}) = 0.996$ whereas $p(\ell = \text{6-cell}|t = \text{6-cell}) = 0.907$. Using the measured $p(\ell|t)$, we then trained the network with 341 embryos labeled at 111,107 times, along with a validation set of 73 embryos labeled at 23,381 times for early stopping [43]. Finally, we apply dynamic programming [11] to the predicted probabilities to find the most-likely non-decreasing trajectory, ignoring images labeled as empty or degenerate (Figure 1.4, center).

Cell Object Instance Segmentation For the images identified by the stage classifier as having 1–8 cells, we next perform object instance segmentation on each cell in the image. We train a network with the Mask R-CNN architecture [46] and a ResNet50 backbone [47], using 102 embryos labeled at 16,284 times with 8 or fewer cells; we also use a validation set of 31 embryos labeled at 4,487 times for early stopping [43]. Our instance segmentation model takes as input a single-focus image cropped by the zona classifier and resized to 500×500. The segmentation model then predicts a bounding box, mask, and confidence score for each detected cell candidate (Figure 1.5, left). Both the ground-truth labels and the predicted masks overlap significantly when the embryo has 4 – 8 cells (Figure 1.5, center). We train the cell segmentation model to minimize the sum of losses of region proposal module, bounding box module, and mask generation module. We produce a final prediction by running our segmentation model on the three central focal planes; we merge candidates

found across focal planes by using the one with the highest confidence score.

Pronucleus Object Instance Segmentation Finally, in the images identified as 1-cell by the stage classifier, we detect the presence of pronuclei. To do so, we train another object instance segmentation network with the Mask R-CNN architecture[46] and a ResNet50 backbone[47]. We use a training set of 151 embryos labeled at 9,250 times during the 1-cell stage, with a validation set of 33 embryos labeled at 1,982 times for early stopping[43]. Pronuclei are only visible during a portion of the 1-cell stage; correspondingly, about 38% of the training images contain 0, 6% contain 1, and 54% contain 2 pronuclei. The pronuclei detector takes as input a single image, cropped by the zona classifier and resized to 500×500, and it predicts a bounding box, mask, and confidence score for each detected candidate (Figure 1.6, left). We run the pronuclei detector on the three middle focal planes and merge candidates found in multiple focal planes by using the one with the highest confidence score.

1.4 Results

We now evaluate our pipeline's performance, demonstrating the effect of each design choice in the models with ablation studies.

Zona Pellucida Segmentation Our zona pellucida classifier nearly optimally segments the test set images, taken from 36 embryos at 576 times. The FCN correctly labels image pixels 96.7% of the time, with per-class accuracies between 93-99% (Figure 1.2, right). The misclassified pixels arise mostly at region boundaries, roughly corresponding to the few-pixel human labeling inprecision at region boundaries.

Fragmentation Scoring The network predicts a score with a mean-absolute deviation of 0.45 from the test labels on the fragmentation test set of 216 embryos labeled at 3,652 times (Figure 1.3, right). When distinguishing between low- (< 1.5) and high- (≥ 1.5) fragmentation, the network and the test labels agree 88.9% of the time. Focus averaging and cropping to a region-of-interest each

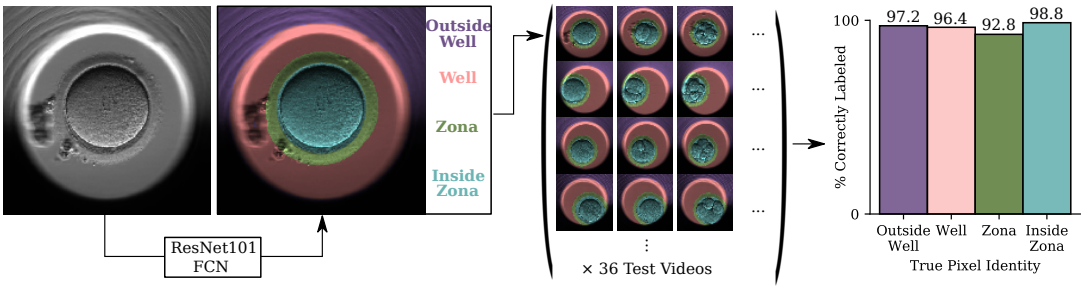

Figure 1.2: Zona pellucida segmentation classifier. The zona classifier (ResNet101 FCN) performs semantic segmentation on the input image, predicting four class probabilities for each pixel (colored as purple: outside well, pink: inside well, green: zona pellucida, cyan: inside zona). Middle: 12 representative segmentations of 3 embryos from the test set. Right: the per-pixel accuracies of the classifier on each class in the test set.

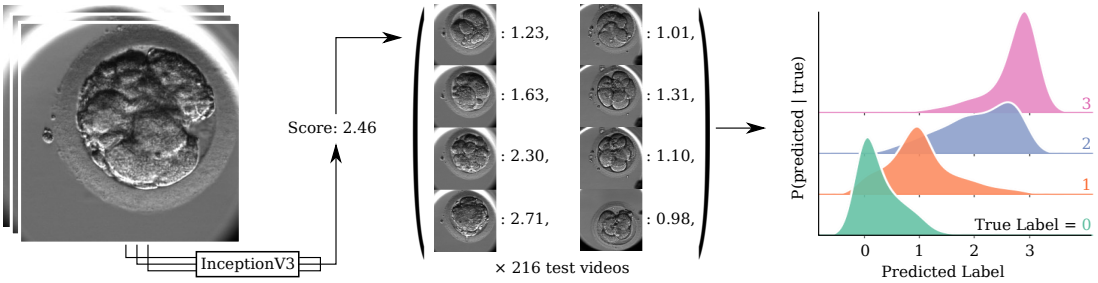

Figure 1.3: Degree of fragmentation classifier. Left: The fragmentation network (InceptionV3 architecture) scores embryos with a real number from 0 – 3; the image at left is scored as a fragmentation of 2.46. Middle: 8 representative fragmentation scores on the test set, shown as image: score pairs. Right: The distribution of the network's prediction given the ground-truth label on the test set. The green distribution corresponds to images with a ground-truth label of 0; orange those labeled as 1; blue, 2; pink, 3.

provide a 1–1.5% boost to the network's accuracy (table 1.1).

We suspect that much of the fragmentation network's error comes from imprecise human labeling of the train and test sets, due to difficulties in distinguishing fragments from small cells and due to grouping the continuous fragmentation score into discrete bins. To evaluate the human labeling accuracy, two annotators label the fragmentation test set in duplicate and compare their results. The two annotators have a mean-absolute deviation of 0.37 and are 88.9% consistent in distinguishing low- from high- fragmentation embryos. Thus, the fragmentation CNN performs nearly optimally in light of the labeling inaccuracies.

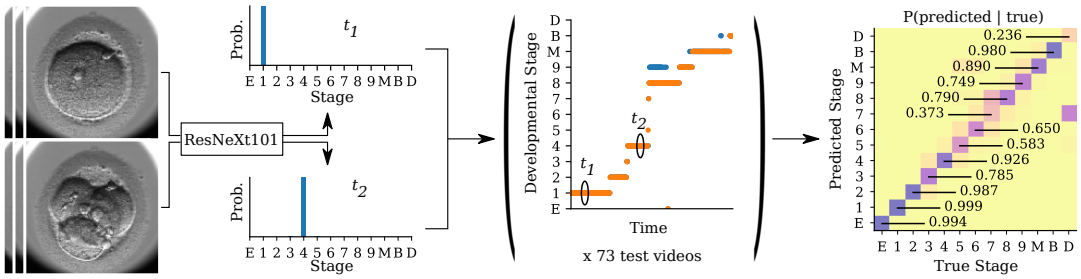

Figure 1.4: Developmental stage classifier. Left: The stage classification CNN (ResNeXt101) predicts a per-class probability for each image; the two bar plots show the predicted probabilities for the two images. Middle: We use dynamic programming to find the most-likely non-decreasing trajectory (orange); the two circled times t_1 and t_2 correspond to the predictions at left. Right: The distribution of predictions given the true labels, as measured from the test set.

Stage Classification The stage classifier predicts the developmental stage with a 87.9% accuracy on the test set, consisting of 73 embryos labeled at 23,850 times (Figure 1.4, right). The network's accuracy is high but lower than the human labeling accuracy on the test set (94.6%). Both the soft-loss and the dynamic programming each improve the predictions by 2% (table 1.1). The stage classifier struggles to distinguish the developmental stage when there are between 5 and 8 cells (66.9% accuracy for these classes). In contrast, the stage classifier does exceedingly well on images with 1-cell (99.9%), 2-cells (98.7%), empty wells (99.4%), or blastocysts (98.0%; Figure 1.4, right). Despite measuring significantly more developmental stages, our stage classifier outperforms previous cell counting networks developed for human embryos [52,62].

Cell Object Instance Segmentation We measure the accuracy of the cell instance segmentation network using mean-average precision (mAP)[70], a standard metric for object instance segmentation tasks. Our network predicts cell masks with a mAP of 0.737 on the test set, consisting of 31 embryos labeled at 4,953 times. The model identifies cells with a precision of 82.8% and a recall of 88.4%, similar to results from other work on fewer images[89]. For correctly-identified candidates, the predicted cell area is within 17% of the true cell area 90% of the time (Figure 1.5, right); much of this error arises when cells strongly overlap late in the cleavage stage. Cropping to a region-of-interest

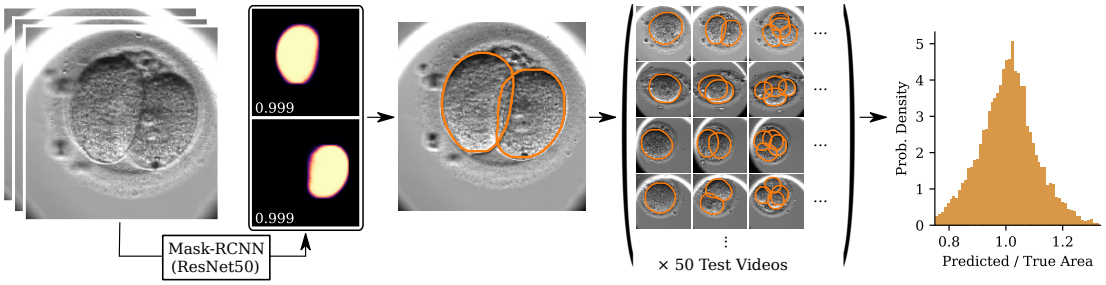

Figure 1.5: Blastomere contour detection network. The cell detection network (Mask-RCNN, ResNet50 backbone) takes an image (left) and proposes candidates as a combined object mask and confidence score from 0–1 (second from left). Center: The boundaries of the object mask represented as the cell's contours (orange, center). Second from left: 12 cell instance segmentations for 4 embryos from the test set (shown as orange contours overlaid on the original image). Right: Histogram of the ratio of predicted to true areas for correctly identified cells in the test set.

Table 1.1: Ablation study of computer vision networks. Effect of design choices for the 4 more difficult tasks, illustrated by removing one modification to the network at a time. Test-set scores are in percent correctly classified (stage, fragmentation) and mean-average precision (blastomere, pronuclei). The best scores are boldfaced.

Setting	Fragmentation (%)	Stage (%)	Blastomere (mAP)	Pronuclei (mAP)
Full Setting	**88.9**	**87.8**	0.737	**0.680**
Single Focus	87.8	84.8	**0.739**	0.668
No ROI from Zona	87.4	84.9	0.733	0.666
Using 50% Training Data	87.7	85.3	0.494	0.656
No Soft Loss	–	85.3	–	–
No Dynamic Programming	–	86.0	–	–

provides a marginal improvement to the network's accuracy (table 1.1).

Pronucleus Object Instance Segmentation The pronuclei segmentation network predicts masks with a mAP of 0.680 on the test set of 33 embryos labeled at 2,090 times. The network identifies pronuclei with a precision of 81.4% and a recall of 88.2%. Much of the false positive detections are from vacuoles inside the 1-cell embryo, which look similar to pronuclei. For correctly-identified candidates, the predicted pronuclei area is within 16% of the true pronuclei area 90% of the time (Figure 1.6, right). Averaging across focal planes and cropping to a region-of-interest improves the mAP by 1% (table 1.1).

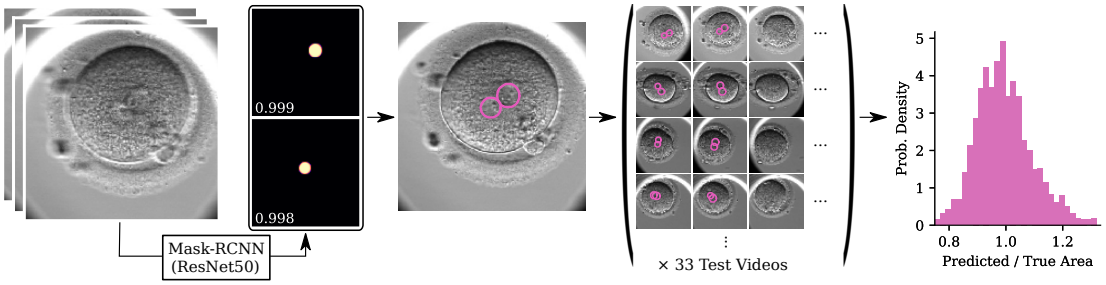

Figure 1.6: Pronuclei detection network. The pronuclei detection network (Mask-RCNN, ResNet50 backbone) takes an image (left) and proposes candidates as a combined object mask and confidence score from 0–1 (second from left). Center: The boundaries of the object mask represented as the pronuclei's contours (magenta, center). Second from left: 12 pronuclei instance segmentations for 4 embryos from the test set (shown as magenta contours overlaid on the original image); the leftmost images illustrate true negatives after the pronuclei have faded. Right: Histogram of the ratio of predicted to true areas for correctly identified pronuclei in the test set.

1.5 Conclusions

Our five networks greatly speed up the grading of embryos: running all five networks on a 300-image, five-day movie takes 6 minutes on a GTX Titan X. In the future, we plan to make this pipeline even faster, by combining all five networks with multi-task learning[30]. Since we measure many of the features used in clinical IVF, our networks have the potential to reduce the time to grade embryos in clinical IVF, without sacrificing interpretability. Equally as important, the automatic, high-quality data from our networks will enable better retrospective chart studies for IVF, improving IVF by informing better clinical practice.

2

BlastAssist: a Deep Learning Pipeline to Measure Interpretable Features of Human Embryos

This chapter is adapted from the manuscript of BlastAssist: a deep learning pipeline to measure interpretable features of human embryos[120].

2.1 Introduction

In vitro fertilization (IVF) has revolutionized the treatment of human infertility and more than 3 million IVF cycles are now performed each year worldwide[1]. Despite the ubiquity of IVF treatments, their success rate remains relatively low: only $\sim 35\%$ of cycles result in live birth in the US[21], 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[81]. Therefore, clinical standards strongly recommend single embryo transfer (SET)[65], which greatly increases the need to accurately evaluate and select embryos.

Manual morphological grading is still the most widely-used method for embryo evaluation[77,37]. Pre-implantation Genetic Testing for Aneuploidies (PGT-A) is increasingly used for embryo selection, though it is invasive and its efficacy is controversial[66,84,76]. Time-lapse microscopy (TLM) incubators have been adopted in many clinics to culture embryos while collecting continuous movies of pre-implantation development[35,9,55]. TLM systems provide significantly more information than traditional manual morphological grading[19,7,20]. 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[4], 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[51], 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[16,102,24] and developing to fetal heartbeats[114,108]. However, such "black-box" approaches have a number of disadvantages[2,96,3]: 1) 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; 2) there are numerous confounders, including patient age, body mass index (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 3) 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[64] 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[52,62,74], blastocyst segmentation of inner cell mass (ICM) and trophectoderm (TE)[90,91,121], and blastocyst grading[61,54]. In addition, automated measurements might be timesaving and assist em-

bryologists, 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 [103,101].

Here, we built off our prior works [63,50,72] and developed BlastAssist, a holistic pipeline to measure a comprehensive set of interpretable features that are clinically and biologically relevant, including: 1) fertilization status, 2) cell symmetry, 3) degree of fragmentation, 4) developmental timing, and 5) 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 1) manual annotations from a panel of human experts in idealized situations, 2) annotations previously made by embryologists during the course of routine IVF treatments, 3) implantation success rate of SET cycles, and 4) the likelihood that the transferred embryo would result in live birth.

2.2 Materials and methods

Study design and dataset

The EmbryoScope^{TM} dataset was collected from the IVF unit of Sourasky Medical Center in Tel Aviv, Israel from treatment cycles performed between 2012-2017. EmbryoScope^{TM} is the most widely-used TLM system for IVF [35]. It utilizes Hoffman modulation contrast (HMC) microscopy [48]. The dataset consists of 32 939 embryos, imaged every 20 minutes at 7 different focal planes up to the first five 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: 1) electronic health record (EHR) data, 2) treatment information, such

as fertilization method, hormone dosage, and culture media, 3) embryo grading and annotations, such as developmental timing, degree of fragmentation on day 2 and day 3, PN count and fade time, and 4) treatment outcomes, such as beta-human chorionic gonadotropin (b-HCG) and live births. These clinical annotations were used to compare clinic and network measurements at 4 different levels: Comparison of BlastAssist measurements to 1) measurements performed by human expert(s) in ideal situations, 2) clinical measurements during routine treatments, 3) implantation outcomes for SET cycles, and 4) live birth results for cycles where up to 4 embryos were transferred.

Metrics

In additions 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 [17] 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 hours 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)[41,12] 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[78] 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[64] 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[105] was performed on each feature to evaluate the statistical significance of the slope of the fitting.

2.3 Results

BlastAssist pipeline

In our previous works[63,50,72], we trained five individual neural networks: 1) ZP segmentation, 2) developmental stage classifier, 3) degree of fragmentation classifier, 4) PN detector, and 5) blastomere detector to measure comprehensive metrics during pre-implantation development. In the current study we trained an additional network: 6) blastocyst segmentation (see Supplementary Data File S1) (Supplementary Table A.1, A.2).

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. 2.1, Supplementary Fig. A.1, A.2, A.3, and A.4).

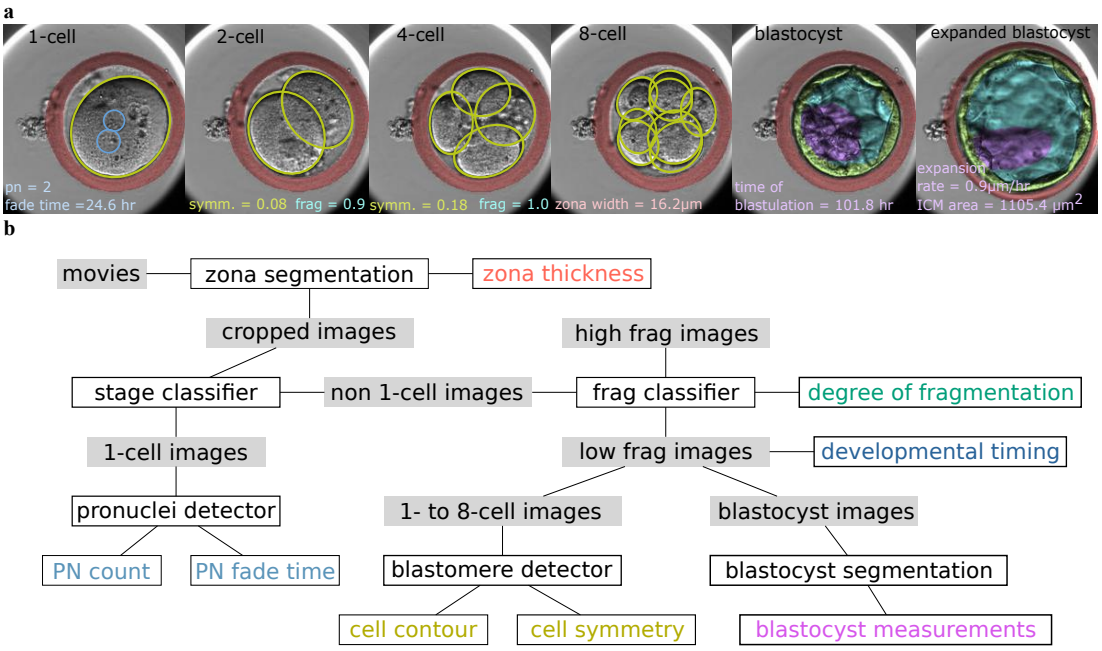

Figure 2.1: Overview of the BlastAssist pipeline of networks. (**A**) An example of BlastAssist outputs on a time-lapse video of an embryo with measurements highlighted. Note: the stage classifier currently do not differentiate between blastocyst and expanded blastocyst, the label of "expanded blastocyst" 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 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 classifier, where we obtain the degree of fragmentation. Only the images with low fragmentation (< 1.5) will be evaluated further. Developmental timing measurement 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.

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 A.3 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[63]. 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 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.

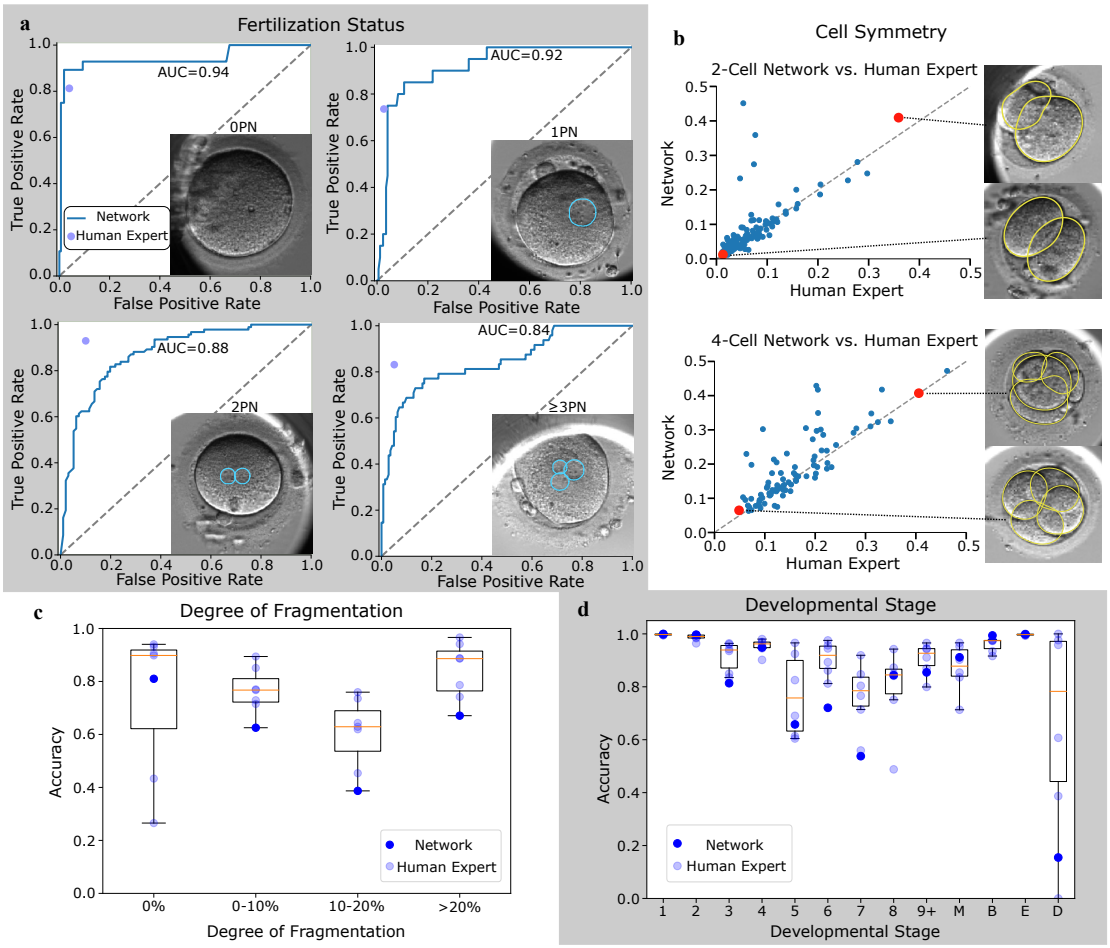

Figure 2.2: Comparison of BlastAsisst outputs to human experts. (**A**) ROC (receiver operating characteristic) curves of network performance in blue on 0PN (pronuclei) (upper left), 1PN (upper right), 2PN (lower left), and $\geq$ 3PN (lower right) embryos with comparison to human expert performance 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.

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. 2.2A). We also measured the human expert's true positive and false positive rate of labeling embryos as having 0, 1, 2, or $\geq$ 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 $\geq$ 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. 2.2B). 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 [83]. 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 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. 2.2C).

For developmental stage, five human experts each labeled the test set, where each labeler anno-

tated 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. 2.2D). 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. A.5) and comparison to clinical annotations when available (Table 2.1). 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. 2.3).

For PN count, the network and embryologists strongly agree, with strong correlation of between the two measurements ($acc = 79.6\%, r = 0.683$). They mostly agree on 1PN and 2PN embryos, with small disagreements on 0PN and $\geq$ 3PN embryos (Fig. 2.3A). The small discrepancies on number of PN mostly come from these following cases: 1) an air bubble or other impurities obstructing the view of the microscope, 2) mis-identifying 2PN with vacuole as $\geq$ 3PN, 3) 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 mistaken 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. 2.3B). 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. 2.3C). 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. 2.3D).

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

Table 2.1. Distributions of EmbryoScope^{TM} dataset measured by BlastAssist and by clinical annotations.

	the entire dataset (32,939 embryos)	dataset where clinical annotations are available (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%
$\geq$ 3PN (abnormal)	6.9%	5.4%	0.3%
PN fade time (hr)	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 (hr)	30.8 ± 13.0	28.6 ± 7.2	28.7 ± 5.3
4-cell time (hr)	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 (hr)	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 ($\mu m/hr$)	1.6 ± 0.7	1.7 ± 0.7	-

Table 2.1: Data are presented in mean ± SD or percentage. ICM: inner cell mass; PN: pronuclei; ZP: zona pellucida.

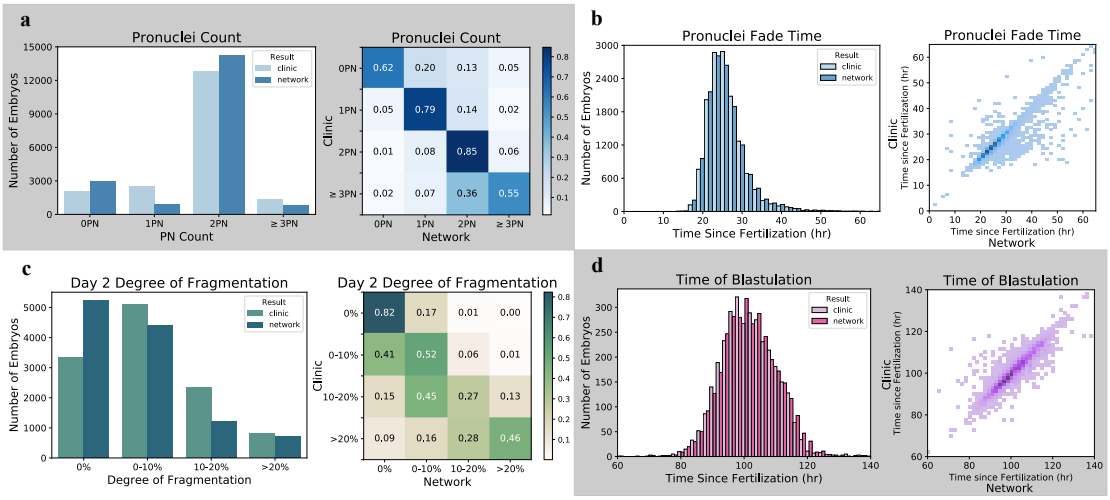

Figure 2.3: Comparison of BlastAssist outputs to clinical annotations. (**A**) Comparison of pronuclei (PN) count measurements between clinic and network, with overall number of embryos per class (left) and confusion matrix of measurements per embryo (right). $n = 18\,922$ embryos. (**B**) Comparison of PN fade time measurements between clinic and network, with overall number of embryos per class (left) and confusion matrix of measurements per embryo (right). $n = 13\,781$ 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). $n = 11\,582$ embryos. (**D**) Comparison of time of blastulation measurements between clinic and network, with overall number of embryos per class (left) and confusion matrix of measurements per embryo (right). $n = 3266$ embryos.

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. 2.4). 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. 2.4A). For developmental timing, delayed cleavage into 2-cell stage is negatively correlated with implantation success rate, with embryos undergoing the first cleavage at $< 24hr$ presenting the highest chance for implantation ($p < 0.01$) (Fig. 2.4B). 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. 2.4C). When analyzing blastocysts, our results show that embryos with higher blastocyst expansion rate ($> 1.5 \mu m/hr$) have higher implantation success rate ($p < 0.2$) (Fig. 2.4D). ICM size of blastocysts also show strong correlation with implantation success rate, with blastocysts presenting ICM size over $1000 \mu m^2$ demonstrating higher implantation rates ($p < 0.5$) (Fig. 2.4E). Amongst the features, developmental timing and cell symmetry are significantly 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).

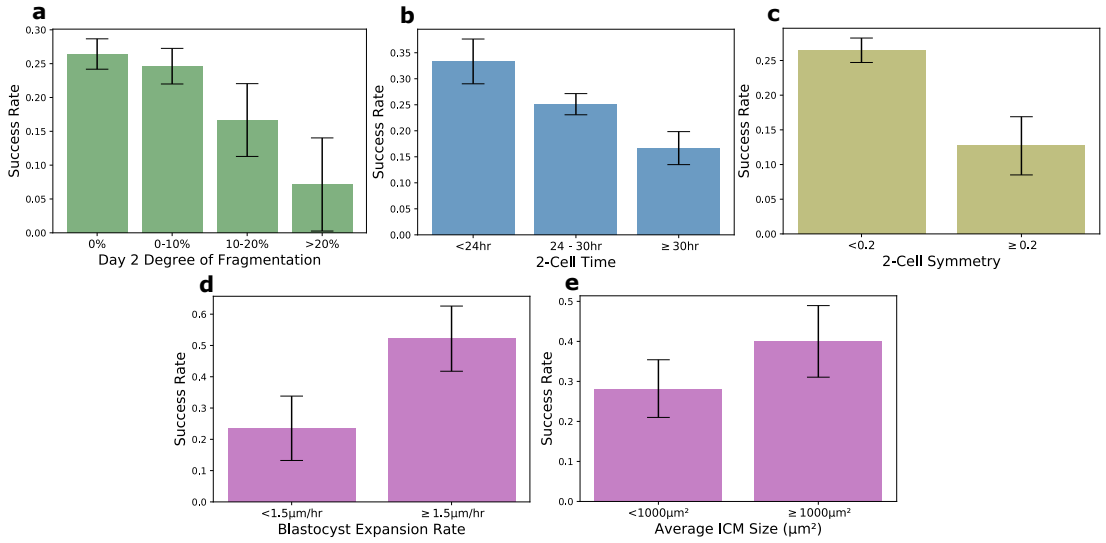

Figure 2.4: Associations of BlastAssist outputs to clinical implantation results. 723 single embryo transfer (SET) cycles were included. (**A**) Implantation success rate for embryos with 0% ($n = 386$), 0-10% ($n = 268$), 10-20% ($n = 48$), and $> 20\%$ ($n = 14$) fragmentation on day 2 post fertilization. $\chi^2 = 4.60, p < 0.3$. (**B**) Implantation success rate for embryos that developed to 2-cell stage in < 24 hours ($n = 120$), $24 - 30$ hours ($n = 454$), and ≥ 30 hours ($n = 138$) 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 $< 1.5\mu m/hr$ ($n = 17$), and $\geq 1.5\mu m/hr$ ($n = 23$) during blastocyst stage. $\chi^2 = 2.25, p < 0.2$. (**E**) Implantation success rate for embryos with inner cell mass (ICM) size $< 1000\mu m^2$ ($n = 39$), and $\geq 1000\mu m^2$ ($n = 30$) during blastocyst stage. $\chi^2 = 0.60, p < 0.5$.

Associations of BlastAssist outputs to clinical live birth results

SET cycles give us the most accurate information on individual embryo's clinical outcome. However, SET cycles have usually been limited to younger patients and patients with good prognosis[113,31], resulting in a relatively small sample size, even from our large clinical dataset. Therefore, we used a previously developed Bayesian inference approach[64] 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. 2.5). We found the correlations between the probability of an embryo resulting in a live birth and individual BlastAssist measurements (Fig. 2.5, Supplementary Fig. A.6).

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. 2.5A). 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. 2.5B). 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. 2.5C). 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. 2.5D). 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. 2.5E). 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. 2.5F).

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 ($n = 407$).

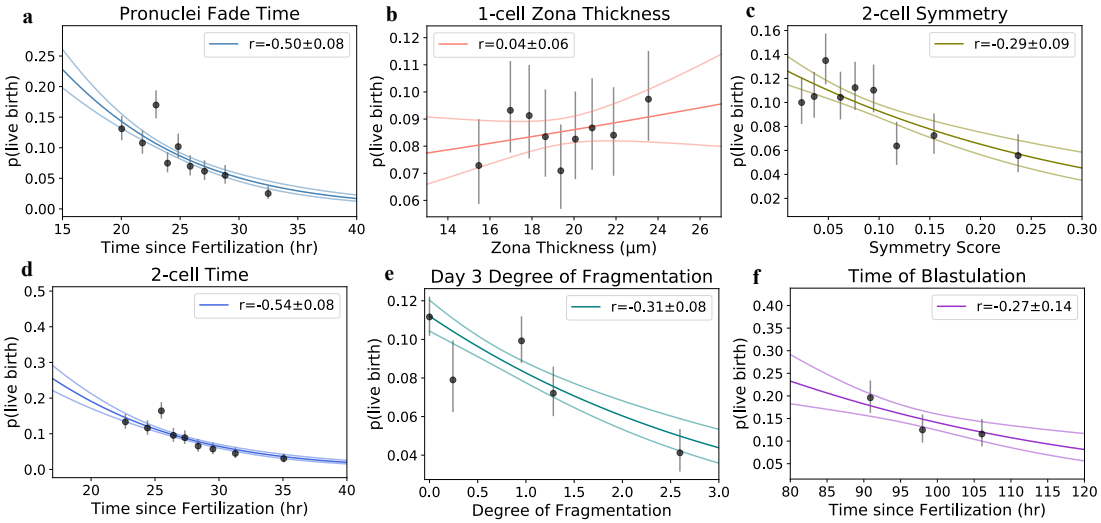

Figure 2.5: Associations of BlastAssist outputs to clinical live birth outcomes. 3421 embryos (1801 cycles) were included in the association study. Estimated probability of an embryo resulting in a live birth as a function of (**A**) pronuclei (PN) fade time ($n = 3321, t = -6.25, p < 5 \times 10^{-10}$). (**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 $p(livebirth)$ and the variables are plotted in each figure. (See Supplementary Data File S1: Correlations to live birth)

Results from additional features are included in Supplementary Fig. A.6.

2.4 Discussion

In this study, we have built off of our prior works[63,50,72] 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 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. 2.5A and B), while previous studies have given conflicting indications regarding the extent to which PN fade time [10,57,27,68] and ZP thickness [69,58,44,42] 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. 2.5C, D, and E), which is consistent with findings of previous studies regarding these variables [88,123,117,99,68,32,67]. Such associations should not form the sole basis of embryo selection algo-

rithms since they do not account for extensive confounding factors[3,96,2]. However, incorporating BlastAssist outputs into mechanistic mathematical models, which explicitly account for patient factors and clinical practice[64], 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.

3

Deciphering IVF: Understanding Causal Relationships in Clinical Features

3.1 Introduction

The clinical IVF environment is extremely complex, with many variables at play, from patient data, treatment data, to embryo data, and IVF outcomes, and many confounding variables. Most clinical IVF studies have used classical statistics to find correlations between variables, hoping to gain understanding of this complex process. However, classical statistics does not take prior knowledge into account, which often leads to the correlation and causation fallacy that we are all familiar with. Classical statistics is especially not suitable to study these complex systems such as the clinical IVF process. In addition to many confounding variables and the complexity of the problem, most fundamentally, typical correlation studies are aimed 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?

Bayesian statistics uses Bayes' theorem to update probabilities for a hypothesis based on new data or information. It is a statistical approach that uses prior knowledge or beliefs, along with new data, to estimate the probability of a hypothesis [111]. Recognizing the limitations of classical statistics, many researchers have adapted Bayesian statistics to build causal models of clinical IVF [64,107,110,79].

In this study, our goal is to build a comprehensive model of the complex IVF process using Bayesian statistics, more specifically, probabilistic graphical models (PGM) to study and understand causal relationships between variables in clinical IVF. PGMs are graphical representations of complex conditional dependencies, which can provide us insight into the model interpretation [59,116]. We utilize the electronic health record (EHR) data from three IVF labs: IVF unit of Lis Maternity Hospital at Sourasky Medical Center in Tel Aviv, Israel, IVF lab of Brigham and Women's Hospital in Boston, MA, USA, and IVF group at Weill Cornell Medicine Center for Reproductive Medicine

and Infertility in New York, NY, USA, in addition to the BlastAssist pipeline output from Aim 2. Due to the high complexity of the IVF process and the present of data missingness, we construct the PGMs in three parts, 1) ovarian stimulation in IVF, the process where embryologists use hormones to stimulate the ovaries to produce multiple eggs instead of just one in a natural mentral cycle. 2) fertilization, where eggs are either co-cultured or injected with sperm to produce a zygote. 3) pre-implantation embryo development, where fertilized embryos are cultured *in vitro*.

With these PGMs, we hope to decipher the complex relationships of variables in clinical IVF and gain insight into the IVF process. By uncovering these hidden patterns and correlations, we may be able to create personalized treatment plans tailored to each individual patient's unique needs. This could lead to higher success rates and improved outcomes for IVF. Moreover, our study may help identify potential predictors of ovarian stimulation, successful fertilization, and embryo development. By pin-pointing the key factors that influence these processes, we can potentially further optimize the IVF cycle and minimize unnecessary steps.

3.2 Materials and methods

Dataset

The EHR data are collected from three IVF labs: the IVF unit of Lis Maternity Hospital at Sourasky Medical Center in Tel Aviv, Israel, the IVF lab of Brigham and Women's Hospital in Boston, MA, USA, and IVF group at Weill Cornell Medicine Center for Reproductive Medicine and Infertility in New York, NY, USA. The embryo EHR data are collected from the IVF unit of Lis Maternity Hospital at Sourasky Medical Center in Tel Aviv, Israel. The morphokinetic features of pre-implantation of embryo development are collected from the BlastAssist pipeline applied to EmbryoScope^{TM} movies from Lis Maternity Hospital at Sourasky Medical Center in Tel Aviv, Israel (Aim 2).

Table 3.1. Dataset sizes for ovarian stimulation PGMs.

Clinic	LMH	BWH	WCM
entire cycle EHR dataset	**51,157**	**10,059**	**1,810**
Eggs, E2, MII	13,869	7,153	1,806
Eggs, E2, MII, Age, BMI	5,261	7,128	1,751
Eggs, E2, MII, Age, BMI, FSH, LH	5,191	6,855	1,751

Table 3.1: dataset size in n cycles. Age: maternal age; BMI: body mass index; BWH: Brigham and Women's Hospital; EHR: Electronic Health Record; Eggs: number of oocytes produced; E2: estradiol; FSH: follicle stimulating hormone; LH: luteinizing hormone; LMH: Lis Maternity Hospital; MII: number of metaphase II oocytes; PGM: probabilistic graphical models; WCM: Weill Cornell Medicine

Table 3.2. Dataset sizes for fertilization PGMs.

Clinic	LMH
entire embryo EHR dataset	**256,559**
Age, Eggs, E2, fert method, 2pn	71,989
Age, Eggs, MII, mature, fert method, 2pn	41,376
Age, fert method, 1-cell zona thickness, zygote size, 2pn	18,101

Table 3.2: dataset size in n embryos. Age: maternal age; EHR: Electronic Health Record; Eggs: number of oocytes produced; E2: estradiol; fert: fertilization LMH: Lis Maternity Hospital; mature: whether an individual oocyte is in MII phase. MII: number of metaphase II oocytes; PGM: probabilistic graphical models; pn: pronuclei

To construct each PGM, relevant features are selected based on both prior knowledge and classical statistical correlations between pairs of variables. The sample sizes of sub-datasets used to construct each PGM are listed in Table 3.1, 3.2, and 3.3.

Statistics

Pearson's correlation coefficient (r)[41,12] was used to evaluate classical statistical correlations between pairs of variables.

Table 3.3. Dataset sizes for embryo development PGMs.

Clinic	LMH
entire BlastAssist pipeline dataset	**32,939**
Age, pnf, 2-cell time, 4-cell time, 2-cell symm, 4-cell symm, d2 frag, d3 frag	15,911
Age, pnf, 2-cell time, 4-cell time, 8-cell time, morula time, blast time, 2-cell symm, 4-cell symm, d2 frag, d3 frag	7,168

Table 3.3: dataset size in n embryos. Age: maternal age; blast: blastocyst; d2, d3: day 2, day 3; EHR: Electronic Health Record; frag: degree of fragmentation; LMH: Lis Maternity Hospital; PGM: probabilistic graphical models; pnf: pronuclei fading; symm: symmetry.

Bayes's theorem was used to evaluate conditional correlation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.1}$$

, where $P(A|B)$ is the conditional probability of A given B, or the posterior probability of A given B. $P(A)$ is the probability of A, or the prior. $P(B|A)$ is the conditional probability of B given A, or the likelihood. $P(B)$ is the probability of B, or the evidence.

Probabilistic Graphical Model (PGM)

A previously developed method[64] was used to construct PGMs in clinical IVF data. All possible relevant variables (nodes) based on prior knowledge are first checked for classical statistical correlations in pairs. If two variables are correlated with each other, then conditional correlation correlations are checked conditioned on all other variables. If the two variables are no longer correlated conditional on any combinations of other variables, then the two nodes are not connected.

The directed acyclic graphs (DAGs) are then constructed based on prior knowledge, presence of colliders, and whether the DAG fits the data. The conditional correlations between connected nodes are then updated based on the DAGs. The final conditional correlations between nodes are

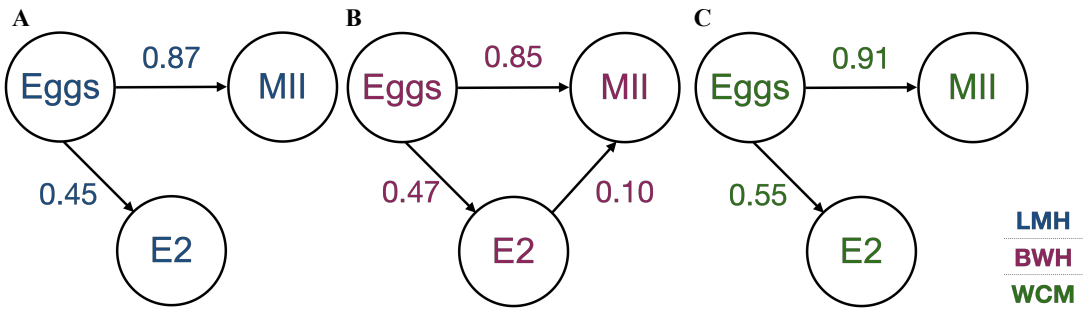

Figure 3.1: Graphical models of ovarian stimulation with variables: Eggs, E2, and MII of (**A**) Lis Maternity Hospital (LMH), (**B**) Brigham and Women's Hospital (BWH), and (**C**) Weill Cornell Medicine (WCM). The weights are correlations between two variables conditional on any incoming nodes.

calculated conditioned on all incoming nodes.

3.3 Results

Ovarian Stimulation in IVF

We first constructed the simplest graphical models within ovarian stimulation, with Eggs, E2, and MII Eggs represent the number of oocytes retrieved from the patient post ovarian stimulation. E2 represents the maximum value of estradiol recorded during ovarian stimulation. In most clinics, E2 is measured throughout the period of ovarian stimulation. As follicle grows in the ovaries, the E2 level also increases[98]. E2 therefore informs embryologists the level of follicle growth and can potentially inform the stimulation protocol. MII represents the number of mature oocytes retrieved. For cycles where intracytoplasmic sperm injection (ICSI) was used as the fertilization method, the oocytes were denuded and the maturation status was checked. Only oocytes in metaphase II (MII) are ready to be fertilized. Sometimes embryologists would attempt to mature the eggs *in vitro*. We developed the models separately for the three clinics (Fig. 3.1). For LMH and WCM, there's the two variables E2 and MII are not correlated when conditioned on Eggs (Fig. 3.1A and C), however they are conditionally correlated for data from BWH (Fig. 3.1B).

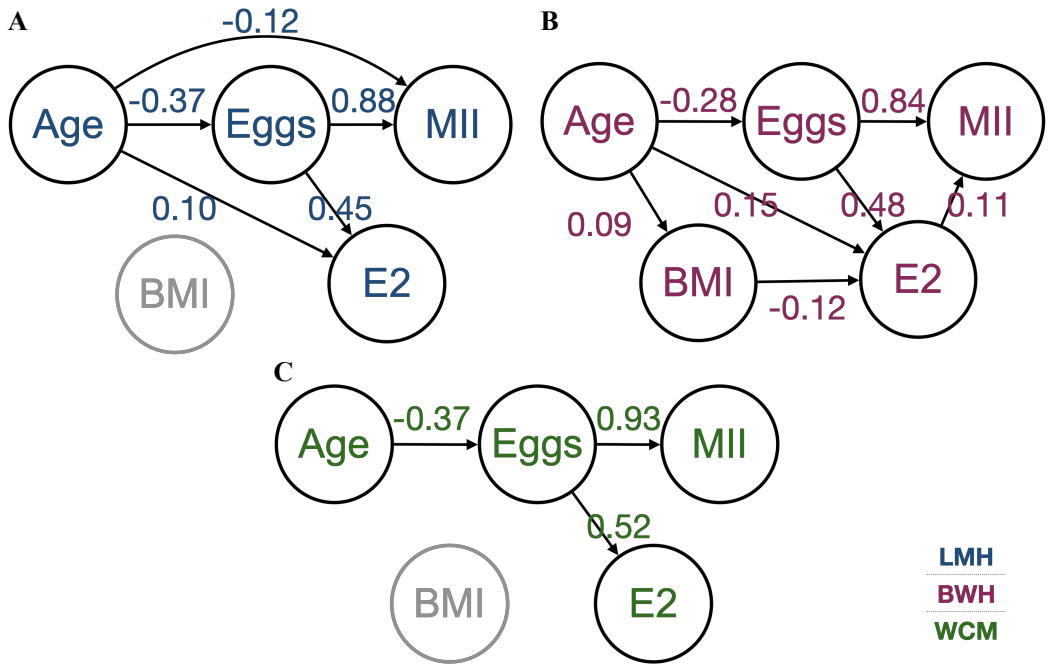

Figure 3.2: Graphical models of ovarian stimulation with variables: Eggs, E2, MII, Age, and BMI of (**A**) Lis Maternity Hospital (LMH), (**B**) Brigham and Women's Hospital (BWH), and (**C**) Weill Cornell Medicine (WCM). The weights are correlations between two variables conditional on any incoming nodes.

We then included Age and BMI in our graphical model. Age represents the maternal age which is defined as the age of the producer of the oocyte (either the patient or egg donor) at the time of oocyte extraction. It is believed to be the most significant factor of many variables in IVF and clinical outcomes [109,119,94]. BMI represents body mass index of the oocyte producer. High BMI is believed to be associated with decrease in IVF success rate [93,100]. From the models we developed separately for the three clinics (Fig. 3.2), we can see that all three clinics' models share similar conditional dependencies. For LMH and WCM, BMI is not connect to any other nodes on the graphical models (Fig. 3.2A and C). What is interesting is that for both BWH and WCM, Age and MII have no correlations when conditioned on Eggs (Fig. 3.2B and C), which may suggest that even though Age is strongly correlated to many variables in clinical IVF, there may not always be conditional dependencies.

To construct the complete model for ovarian stimulation, we then included FSH and LH into the model. FSH represents follicle-stimulating hormone, a hormone produced by the anterior pituitary in response to gonadotropin-releasing hormone (GnRH) from the hypothalamus. It plays an essential role in follicle development. In addition to our body's natural FSH, FSH medication, along with luteinizing hormone (LH), are used during ovarian stimulation to increase egg production and improve egg maturation. LH is another hormone produced in response to GnRH from the hypothalamus. LH surge triggers ovulation, and the development of corpus luteum. During the follicular phase, normal follicular growth is the result of complementary action of FSH and LH [92]. Thus, LH is often administered along with FSH during ovarian stimulation. We broke down the graphical models into three regions, patient variables (Age, BMI), treatment variables (FSH, LH), and response variables (Eggs, E2, MII) and constructed seperate PGMs for each clinic (Fig. 3.3). Both BWH and WCM administer their hormones using FSH and HMG (Human Menopausal Gonadotropin), which is part FSH and part LH. Therefore the models from these two clinics show significant auto correlation between FSH and LH (Fig. 3.3B and C). We see that in the models of all

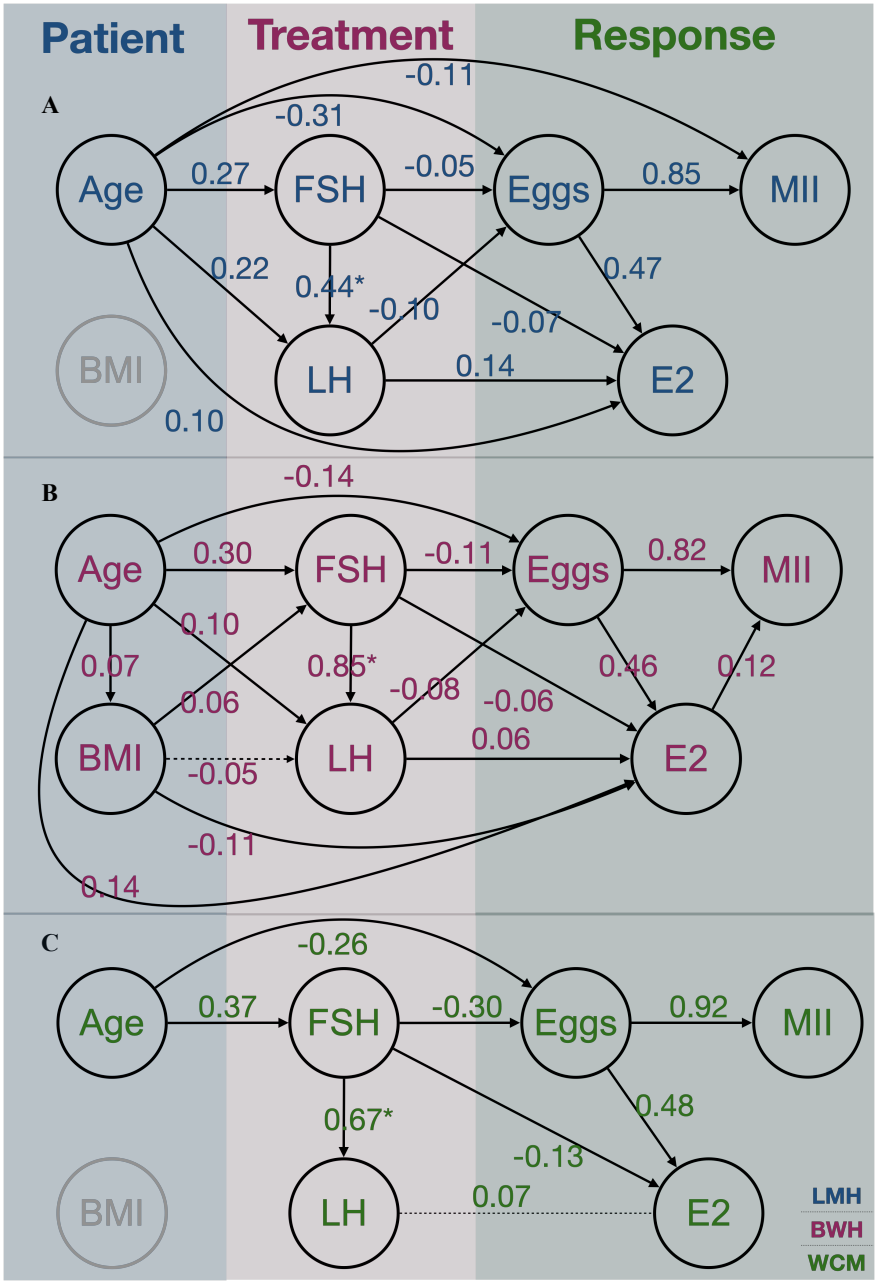

Figure 3.3: Graphical models of ovarian stimulation with variables: Eggs, E2, MII, Age, BMI, FSH, and LH of (**A**) Lis Maternity Hospital (LMH), (**B**) Brigham and Women's Hospital (BWH), and (**C**) Weill Cornell Medicine (WCM). The weights are correlations between two variables conditional on any incoming nodes. The asterisk (*) represent that both direction of the arrow can be correct, and an arbitrary direction is used to construct the DAG.

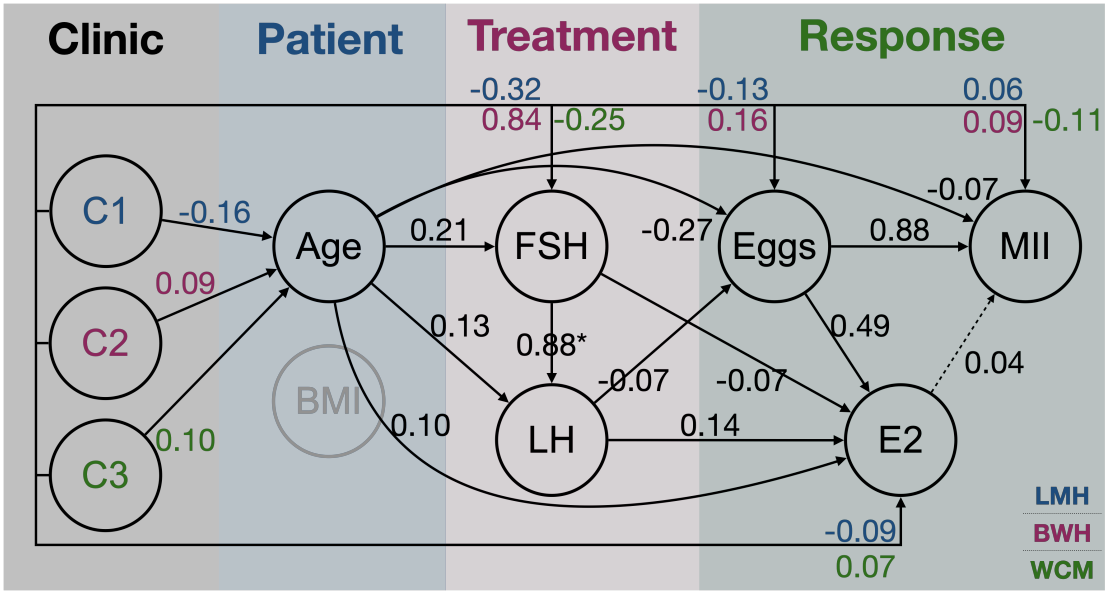

Figure 3.4: Graphical model of ovarian stimulation with variables: Eggs, E2, MII, Age, BMI, FSH, LH, and Clinic of the combined dataset of Lis Maternity Hospital (LMH), Brigham and Women's Hospital (BWH), and Weill Cornell Medicine (WCM). The weights are correlations between two variables conditional on any incoming nodes. The asterisk (*) represent that both direction of the arrow can be correct, and an arbitrary direction is used to construct the DAG.

three clinics, both FSH and LH are negatively associated with Eggs, which may suggest over stimulation. This finding is consistent with other studies suggesting the current standard IVF dosage may be too high[33,6]. It is important to note that WCM's model is the most simplistic (Fig. 3.3C), likely due to its smaller sample size, therefore not being able to detect subtle conditional dependencies.

In order to understand the effect of individual clinics, we then encoded the clinic as a categorical variable using dummy coding[45]. We modeled the clinics as three variables with some random noise (normal distribution with $\bar{x} = 0, \sigma = 10^{-4}$) to avoid the error of dividing by zero. When checking the effect of each clinic in the causal model, we get the combined model of all clinics (Fig. 3.4). When all the data are combined it becomes clear that BMI is not a node in the model. The clinical factor affects the maternal age likely due to the fact that LMH is not allowed to accept patients over the age of 45 due to government regulations. The effect on E2 is likely due to the different proto-

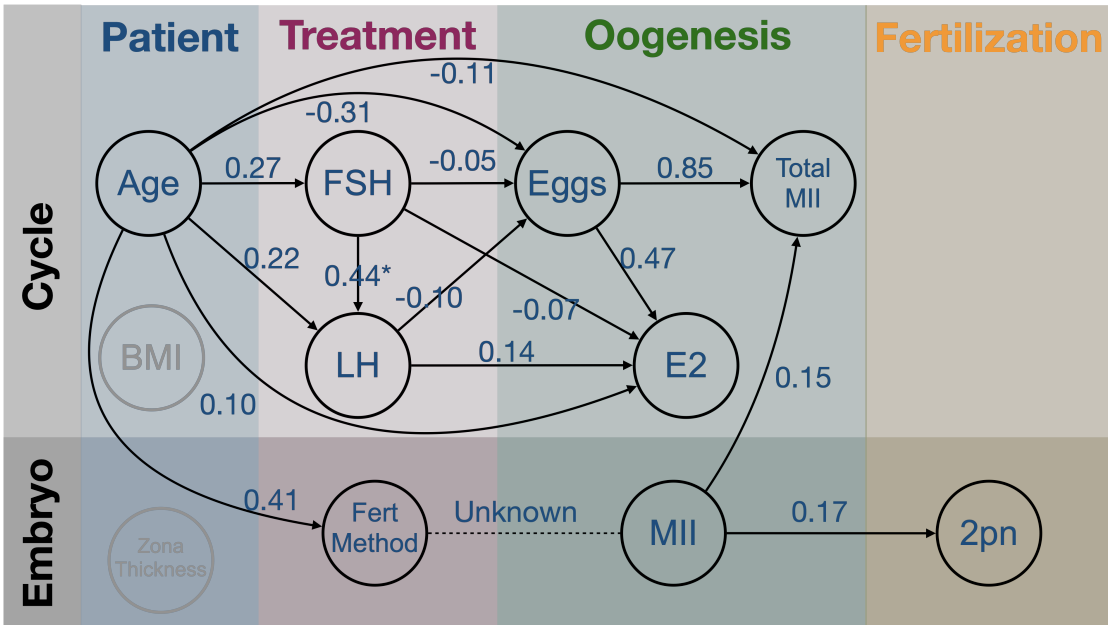

Figure 3.5: Graphical model of fertilization of the dataset from Lis Maternity Hospital (LMH). The weights are correlations between two variables conditional on any incoming nodes. The asterisk (*) represent that both direction of the arrow can be correct, and an arbitrary direction is used to construct the DAG.

cols of E2 measurements and recording. The effect on FSH is likely due to different stimulation protocols, which also affects the subsequent production of eggs and MII eggs.

Fertilization

After constructing a complete model for ovarian stimulation, our next step is to build a graphical model for fertilization (Fig. B.3). Only data from LMH is included due to the fact that we only have morphokinetic data such as zona thickness and zygote size from the BlastAssist pipeline. We first investigated if there is conditional dependencies between Age, Eggs, E2, fertilization method, and 2pn. Fertilization method represents the technique used to fertilize the oocyte, which can either be IVF, where oocytes are co-cultured with sperm, or ICSI, where one sperm is injected directly into the oocyte.

We first excluded any variable that involves maturation status of oocyte, due to the fact that the

maturation status is only known for embryos that undergoes the ICSI procedure. We see correlation between Age and fertilization method, which may suggest that different fertilization method is used based on maternal age. However, we know that ICSI procedure is usually utilized based on male infertility factors, therefore this correlation could be resulted from a confounding factor such as paternal age, which tend to be correlated with maternal age. We see that there is no conditional dependencies between fertilization method and results.

We then investigated the correlations between MII, mature status, and 2pn. We found that an individual oocyte's mature status contributes to both the total number of MII eggs retrieved from the cycle, and the fertilization status of the embryo, which both make intuitive sense. We are unable to investigate whether there's conditional dependencies between fertilization method, the oocyte's mature status, and fertilization status, due to the fact that maturation is only checked for ICSI cycles.

Prior studies have shown association between zona pellucida thickness and fertilization outcome [14,15,75]. However, we found that neither zona pellucida thickness or zygote size play a part in fertilization rate in cycles where both IVF and ICSI fertilization methods were used.

Morphokinetic Features in Pre-implantation Embryo Development

Lastly, we built graphical models of pre-implantation embryo development from morphokinetic features we measured using the BlastAssist pipeline [120]. The variables we included are features typically measured in IVF clinics. The variables include developmental timing information: pn fade time, 2-cell time, 4-cell time, 8-cell time, morula time, and blastulation time; cell symmetry measurements: 2-cell symm, 4-cell symm; degree of fragmentation: d2 frag, d3 frag.

We first built a graphical model of embryo development using morphokinetic features up to the 4-cell stage (Fig. 3.6). We found that up to the 4-cell stage, pn fade time directly contribute to 2-cell time, and 2-cell time directly contribute to 4-cell time, with no conditional dependencies between

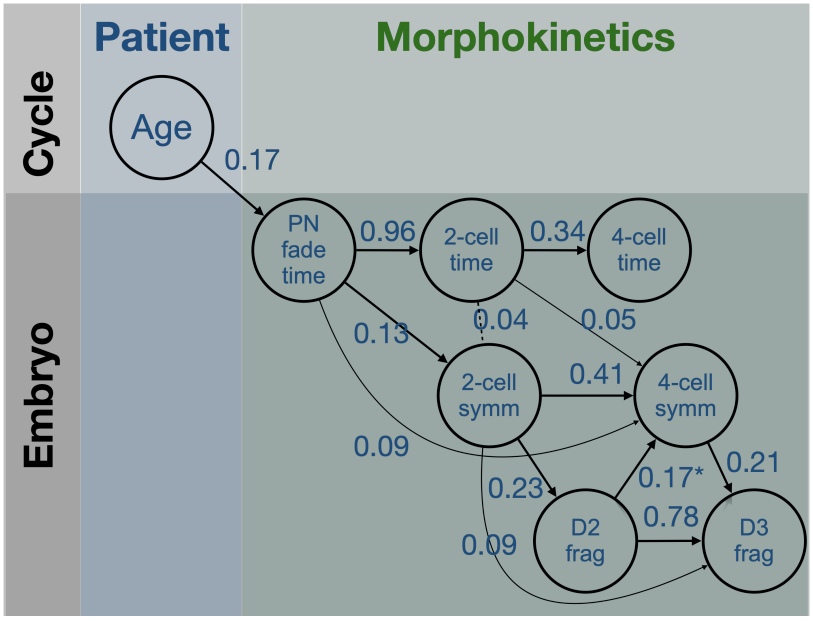

Figure 3.6: Graphical model of embryo development up to 4-cell stage of the BlastAssist pipeline dataset from Lis Maternity Hospital (LMH). The weights are correlations between two variables conditional on any incoming nodes. The asterisk (*) represent that both direction of the arrow can be correct, and an arbitrary direction is used to construct the DAG.

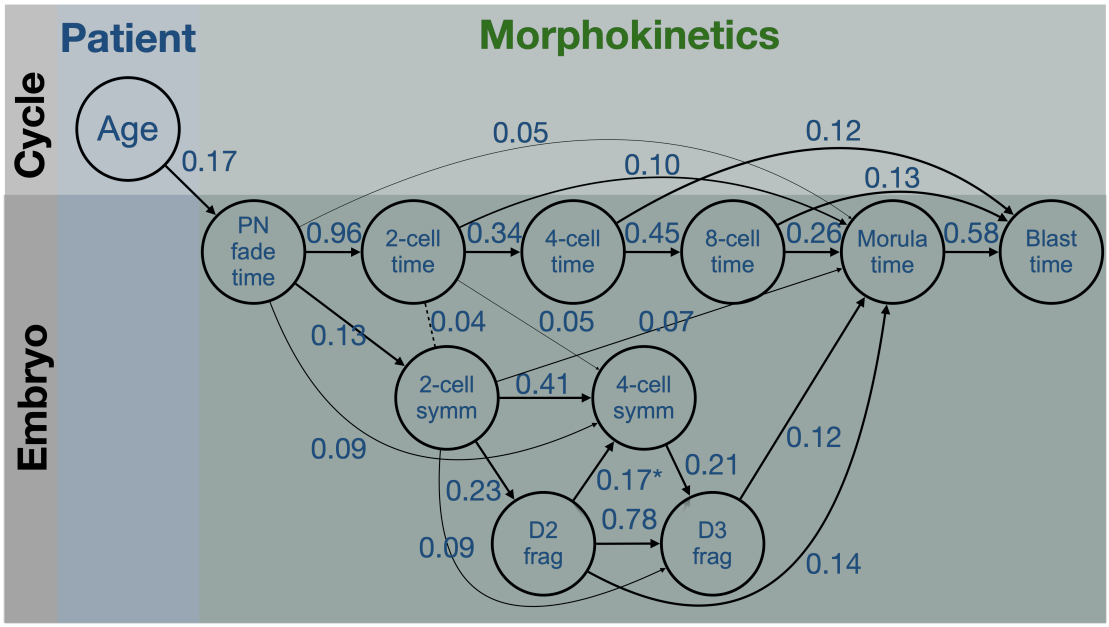

Figure 3.7: Graphical model of embryo development of the BlastAssist pipeline dataset from Lis Maternity Hospital (LMH). The weights are correlations between two variables conditional on any incoming nodes. The asterisk (*) represent that both direction of the arrow can be correct, and an arbitrary direction is used to construct the DAG.

on fade time and 4-cell time. We see relatively strong conditional correlations between cell symmetry measurements and degree of fragmentation measurements.

We then included morphokinetic features up to the blastocyst stage and constructed the graphical model (Fig. 3.7). We see that many early stage morphokinetic features have causal relationship with morula time, suggesting that these features may have causal relationship with morula formation, which is the prior of morula time. And these same features have no effect on blastulation time. On the contrary, two features: 4-cell time and 8-cell time have no effect on morula time, but have causal relationships with blastulation time. This result suggest that morula formation and blastulation may be two processes determined by completely orthogonal factors. It is interesting to point out that in both embryo development models (Fig. 3.6 and 3.6), Age has no conditional dependencies with any downstream morphokinetic features.

3.4 Discussion

In conclusion, this study aimed to construct a comprehensive model of the IVF process using Bayesian statistics, specifically probabilistic graphical models (PGM), to elucidate causal relationships among variables in clinical IVF. Using EHR data from three IVF clinics and BlastAssist pipeline output, we partitioned the construction of PGMs into three parts: ovarian stimulation, fertilization, and pre-implantation embryo development. Through these models, we can potentially reveal the complex causal relationships of variables in clinical IVF, gaining valuable insights into the process. Additionally, findings from this study can be used to direct further clinical research. Overall, the application of Bayesian statistics and PGMs in IVF represents a promising future for advancing our understanding of this complex IVF process and improving clinical outcomes.

4

Conclusion and Future Work

By integrating AI and Bayesian statistics, we have gained significant insights into the highly complex process of clinical IVF. We first improved the accuracy and efficiency of Time-Lapse Microscopy (TLM) image analysis using Computer Vision (CV). Utilizing cutting-edge CV technique, we have automated the extraction of morphokinetic features from TLM movies of human embryos. These features offer a more objective assessment of embryo quality, which can potentially aid embryologists in embryo selection and streamline the IVF image analysis process.

Having successfully developed the CV algorithms in Aim 1, we then integrated the algorithms into BlastAssist, a pipeline that measures comprehensive, qualitative, and clinically relevant features in clinical IVF. To validate our pipeline, we conducted detailed comparisons between BlastAssist pipeline measurements with human experts' manual measurements, annotations provided by embryologists during routine treatments, the outcomes of single embryo transfer (SET) cycles, and the live birth outcomes of transferred embryos. This shows that BlastAssist has the potential as a powerful tool for embryologists to improve the embryo selection process and, consequently, improve clinical success rates.

Utilizing the large dataset generated by the BlastAssist pipeline, in conjunction with the EHR data from three IVF labs, we constructed probabilistic graphical models (PGMs) for the complex IVF process in three parts: ovarian stimulation, fertilization, and embryo development. These graphical models elucidates intricate and sometimes subtle causal relationships of variables in clinical IVF. Additionally, findings from these graphical models can be used to direct further clinical IVF research. Overall, the application of Bayesian statistics and PGMs in IVF represents a promising future for advancing our understanding of this complex process and improving clinical outcomes.

In the future, to further improve upon the results from this project, one could potentially include additional factors such as other hormones administered in the treatment process, such as Gonadotropin hormone-releasing hormon (GnRH) and Progesterone, or other variables measured such as Anti-Müllerian hormone (AMH), a measure of ovarian reserve and endometrium thickness.

Another study conducted in the Needleman lab has elucidated the strong cohort effect of developmental timing, where the average developmental timing of the cohort, instead of the variation from the average, is predictive of fetal heartbeat. Further research on the cohort effect of other features such as cell symmetry and degree of fragmentation may be informative to further improvement of the causal models by updating the models with more prior knowledge. In Aim 2, we have used Bayesian generative models to study the correlation between individual morphokinetic features and the probability of live births. The same method can be used to investigate the causal relationships between these features and IVF outcomes.

A

Supplementary Data of Chapter 2

A.1 Detailed data annotation and dataset sizes for classifier training

For each classifier, a subset of the clinical IVF dataset was used to train and optimize its performance. The number of embryos used for training was doubled until it no longer improved the classifier performance. A specific graphic user interface (GUI) was created for each annotation task.

For ZP segmentation, a subset of 266 embryos (4194 images) were randomly selected and split (86% - 14%) into training and test set. No validation set was needed due to the high accuracy of the classifier. Every 10th image of the embryo was labeled by a human expert. Each pixel of the image was labeled as one of 4 classes: outside the well, inside the well, ZP, and inside ZP.

For PN detector, a subset of 217 embryos were randomly selected and split (70% - 15% - 15%) into training, validation, and test set. Every image of 1-cell stage embryos was labeled by a human expert. The bounding boxes of the PN were labeled and the masks of PN were calculated based on the bounding boxes.

For blastomere detector, a subset of 164 embryos were randomly selected from a list of embryos that were labeled as low fragmentation from clinical annotations, since fragmentation can affect a human expert's ability to accurately determine the cell contour. The dataset was split (60% - 20% - 20%) into training, validation, and test set. Every image up to the 8-cell stage was labeled by a human expert. In each image, the contour of each cell is annotated.

For fragmentation classifier, a subset of 1410 embryos were randomly selected and split (70% - 15% - 15%) into training, validation, and test set. Every 10th image of the post 1-cell stage cleavage embryos (2-cell to before compaction) was labeled by a human expert. Each image was labeled as one of 4 classes: 0%, 0%-10%, 10%-20%, and above 20%.

For developmental stage classifier, a subset of 487 embryos were randomly selected from a list of embryos that were labeled as low fragmentation from clinical annotations, since fragmentation can occlude the cells and prevent a human expert's from accurately counting the number of cells and de-

termine the developmental stage. The dataset was split (70% - 15% - 15%) into training, validation, and test set. Every image of the entire video was labeled by a human expert. Each image was labeled as one of 13 classes: 1 - 9+ cells, morula, blastocyst, empty well, and degenerate.

For blastocyst segmentation, a subset of 134 embryos were randomly selected and split (70% - 10% - 20%) into training, validation, and test set. More embryos were assigned to the test set in order to have enough data to estimate human expert labeling accuracy. Every 5th image in the blastocyst stage was labeled by a human expert. In each image, the contour of the inner and outer boundary of TE, and the outer boundary of the ICM were annotated.

The sizes of training, validation, and test set for each classifier are outlined in Supplementary Table A.1.

A.2 Architectures and performances of classifiers

The architectures and the performances of the networks are reported in Supplementary Table A.2. Metrics to evaluate each classifier's performance were selected based on the classification task. We incorporated multiple metrics recommended for CV image analysis[73] where they apply. For object instance segmentation tasks such as PN and blastomere detection, we utilized mean average precision (mAP) as the metric. For image level multi-class classification such as degree of fragmentation, we utilized overall per-class (4 classes) accuracy, balanced accuracy (BA), and high vs. low (2 classes) fragmentation accuracy. For image level multi-class classification with mixed classes such as developmental stage, we have both physiologically typical stages such as 1-cell, 2-cell, 4-cell, 8-cell, morula, blastocyst, and physiologically atypical or transition stages such as 3-cell, 5-7 cell, 9+ cell, empty well, and degenerate. The classes are unique because they are inbalanced however of a relatively constant proportion and time-dependent. Therefore we utilized both overall per-class accuracy (Supplementary Table A.2) and individual class accuracy (Fig. 2.2D). We found that for both semantic

segmentation tasks, the classifiers' accuracies are high and the error is similar to the uncertainty of the measurements done by human experts, when uncertainty is measured by the same human expert labeling the segmentation task twice. Due to the high accuracies of the classifier, especially considering the limitation of human labeling precision, simple per-pixel accuracy suffices as a metric in these two cases. Additional details regarding training and evaluation of the classifiers from a pure computer vision perspective have previously been reported[63,50,72].

For PN and blastomere detection, we trained separate object instance segmentation networks with Mask R-CNN, aka Region-Based Convolutional Neural Network, architecture[46] and ResNet50 backbone[47]. As for fragmentation scoring and developmental stage classifier, we trained separate CNNs for regression (InceptionV3 backbone[106]) and classification (ResNeXt101 backbone[118]). For ZP and blastocyst segmentation, we trained a semantic segmentation networks with Fully Connected Neural Networks (FCNs) and ResNet101 backbone[47]. To optimize the performance of networks, we utilized techniques such as averaging results from multiple focal planes, proofreading annotations, early stopping[43], training with a soft loss function[63], and smoothing trajectories with dynamic programming[11,34].

A.3 Using BlastAssist pipeline to measure clinical features

Abnormal fertilization

The PN network we trained detects PN on every frame of 1-cell stage embryos, up to 90 frames imaged over 30 hours (as identified by the developmental stage classifier). For each identified pronucleus candidate, the network returns both a mask of the pronucleus and a confidence score. We consider the pronucleus candidates real when the confidence scores are above a certain threshold.

Supplementary Figure A.1 shows the per-image network predictions of a 2PN (Supplementary Fig. A.1A) and a 3PN (Supplementary Fig. A.1B) embryo. The network clearly detects both the

appearance time and the fade time of the PN, in addition to their number. However, the maximum number of PN detected by the network can be dominated by the occasional errors the network makes in detections. Supplementary Figure A.1A shows that the network can incorrectly detect 3PN in a 2PN embryo.

We deal with this difficulty by calculating the number of PN with the approach described later, first by transforming the detection result and then by taking the 95th percentile of the expected number of PN (see Calculating number of pronuclei). Qualitatively, the expected number is a smooth approximation of the number of PN detected in each frame; the 95th percentile is then a robust measure of the maximum number of detected PN.

Cell symmetry

The blastomere network identifies cell contours for each cell up to the 8-cell stage of embryo development. To measure cell symmetry for 2-cell and 4-cell stage embryos, we first use the stage network on every frame in the embryo's movie, giving a prediction for the embryo's stage as a function of time. We then run the cell contour detection on all frames identified as 2-cell and 4-cell stage. The cell contour detection network returns a list of detection candidates. We represent each candidate by a confidence score and its boundary curve. For all images in the 2-cell and 4-cell stage, we identify the two or four detection candidates with the highest confidence score as the cells. Choosing detection candidates based on the stage classifier gives better performance than thresholding based on the detection's confidence score alone.

We measure cell symmetry using the combined results of the developmental stage classification network and the cell detection network as described previously in materials and methods (Supplementary Fig. A.2). Because cells in the embryo move and change their pose, and because the cell detector network is not perfect, the symmetry score fluctuates in time. We calculate the embryo's symmetry score as the time-averaged symmetry for both 2-cell (yellow line in Supplementary Fig.

A.2B) and 4-cell stage (yellow line in Supplementary Fig. A.2C).

Fragmentation

The current standard clinical approach to detect and quantify fragmentation is to manually inspect embryos at one or two time points (on day 2 and/or day 3) during cleavage stage and classify the degree of fragmentation into a few discrete categories. However, degree of fragmentation is a continuous instead of discrete variable. Therefore, attempting to represent a continuous variable in discrete categories can cause mis-classification and mis-representation in borderline cases.

Instead, the network we trained outputs a continuous score (from 0 to 3.5) for degree of fragmentation on each frame of a cleavage stage embryo (Supplementary Fig. A.3A). For comparison with manual labels, the continuous score output from the network can then be transformed into four classes of degree of fragmentation used in clinics: 0%, 0%-10%, 10%-20%, and above 20%.

Developmental timing

The network we trained identifies the developmental stage of the embryo as a function of time. It predicts the probability of the embryo's developmental stage, including from 1-cell to 8-cell, 9+cell, morula, blastocyst, and degenerate, for each image in the movie. We use dynamic programming to find the most-likely, non-decreasing trajectory for the embryo's stage, with empty wells (when the embryo is taken out of the EmbryoScope^{TM}) allowed at any point during the trajectory (Supplementary Fig. A.3B).

Blastocyst segmentation and measurements

The blastocyst network segments images into ZP, ICM, blastocoel, and TE (Supplementary Fig. A.4A). We used the segmentation results to obtain quantitative measurements of blastocysts expansion, including dynamic changes in blastocyst diameter, ICM area, TE and ZP thickness, and relative size of blastocoel. We measure the TE and ZP thickness using the approach described in Materials and methods. Supplementary Figure A.4B shows the results from running the network on an

individual embryo. For this embryo, the blastocyst diameter increased over time at a relatively constant rate at $3.07 \pm 0.04 \mu m/hr$ (Supplementary Fig. A.4C). After the formation of the ICM, its area fluctuates around a constant average at $1730 \pm 330 \mu m^2$ (Supplementary Fig. A.4D). The relative size of its blastocoel has a steep increase in the beginning of its formation, and then increases steadily at a relatively constant rate at $3.2 \pm 0.1\%/hr$ as the blastocyst expands (Supplementary Fig. A.4E). The average thickness of its TE also fluctuates but decreases over time as the blastocyst expand (Supplementary Fig. A.4F). In accordance, the average ZP thickness decreases over time as the blastocyst expands at $-0.62 \pm 0.03 \mu m/hr$, and it reaches a more clearly defined plateau of $3.3 \pm 0.1 \mu m$ after $114 hr$ (Supplementary Fig. A.4G).

A.4 Clinical dataset

We applied the BlastAssist pipeline to a dataset of 67 043 973 images (32 939 time-lapse movies of embryos) (Supplementary Fig. A.5, Supplementary Table A.3).

A.5 Calculating number of pronuclei

The PN detection outputs candidates of PN with a mask and a confidence score. The confidence score is a number between 0 and 1 that is the network's estimate of the probability that the detected candidate is real. Instead of simply thresholding the confidence score, which can be noisy for uncertain predictions, we first re-scale the network's confidence score to a probability that more accurately aligns with the data. We define this rescaled probability for each detection as

$$p = \text{logistic}(a \times \text{logit}(c) + b), \tag{A.1}$$

where a and b are fit parameters, c is the confidence score from PN detection output.

We fit these parameters by performing binomial regression on the true number of PN in a test set of data labeled by human experts. After transforming the PN detection outputs from the integer number of true candidates to a continuous, scalar representation of the number of PN per frame, we next calculate the expected number of PN in each frame, by summing the probability function $\sum p$, measured for each frame in the image. We then smooth the expected number in time with a boxcar kernel of width 3 and take the 95th percentile of the expected counts. We use the 95th percentile as a robust measure of the maximum number of detected PN, as the absolute maximum itself is an unstable estimator, and are often dominated by occasional detection errors. We then take this as a feature into 4 separate logistic regression models which predict the probability that an embryo has 0, 1, 2, or $\geq$ 3PN.

A.6 Correlations to live birth

For live birth results of the 3421 embryos (1801 cycles) where up to four embryos were transferred, we used previously developed method[64] 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.

For each cycle c, where n embryos were transferred and resulted in h live births, the probability of an embryo x resulting in a live birth is $p(h|n, p_x)$, where p_x is the probability of the embryo resulting in live birth given the embryo parameters x, which is what we are looking for. The overall data observed would be the product of the probability of all cycles, assuming that each cycle is independent:

$$p(H|N, P_x) = \prod_c p(h_c|n_c, p_{x_c}). \tag{A.2}$$

p_x can be expressed as $p(h|x)$ or $p(x)$ the probability of an embryo resulting in live birth as a func-

tion of embryo parameters x, assuming that x is deterministic of the probability p, and get:

$$p(H|N,X) = \prod_c p(h_c|n_c, p(x)_c). \tag{A.3}$$

Then, we can fit the overall data observed to Eq. (3) and find the best fit to the data to find $p(x)$. We use two types of fitting models, discrete and continuous. For the discrete model, we bin the value of x and fit the value of p for each interval. For the continuous model, we fit $p(x)$ to a Chebyshev polynomial series, more details of the method in [64] (Supplementary Materials 1.1 Regression). This method also assumes that the probability of embryos implanting is independent in cycles with multiple transfers. It has been shown that such assumption is consistent with clinical data in [64].

Supplementary Table S1. Training, validation, and test set sizes of classifiers.

classifier	training set	validation set	test set
ZP	203 embryos (3,618 images)	-	36 embryos (576 images)
PN	151 embryos (9,250 images)	33 embryos (1,982 images)	33 embryos (2,090 images)
blastomere contour	102 embryos (16,284 images)	31 embryos (4,487 images)	31 embryos (4,953 images)
fragmentation	989 embryos (16,315 images)	205 embryos (3,416 images)	216 embryos (3,652 images)
developmental stage	341 embryos (111,107 images)	73 embryos (23,381 images)	73 embryos (23,850 images)
blastocyst	93 embryos (837 images)	11 embryos (95 images)	30 embryos (265 images)

Table A.1: PN: pronuclei; ZP: zona pellucida.

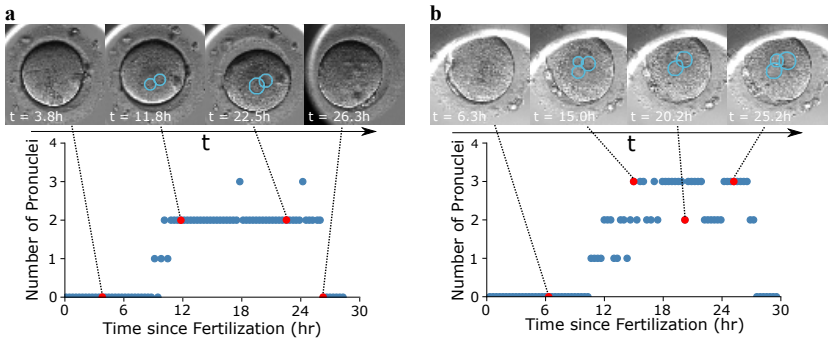

Figure A.1: Detection results of abnormal fertilization. Network prediction in blue dots of number of pronuclei (PN) over the entire 1-cell stage of (**A**) a 2PN and (**B**) a 3PN embryo, with red dots corresponding to sample images in the movie, with PN detection highlighted in blue.

Supplementary Table S2. Architectures and performances of classifiers.

classifier	classification task	architecture	metric	accuracy
PN	object instance segmentation	Mask R-CNN ResNet50	mAP	0.680
blastomere contour	object instance segmentation	Mask R-CNN ResNet50	mAP	0.737
fragmentation	regression	InceptionV3	high vs. low frag per-class acc balanced acc (BA)	88.9% 65.8% 62.1%
developmental stage	classification	ResNeXt101	per-class acc	87.9%
ZP	semantic segmentation	ResNet101	per-pixel acc	96.7%
blastocyst	semantic segmentation	ResNet101	per-pixel acc	88.8%

Table A.2: acc: accuracy; mAP: mean average precision; PN: pronuclei; ZP: zona pellucida.

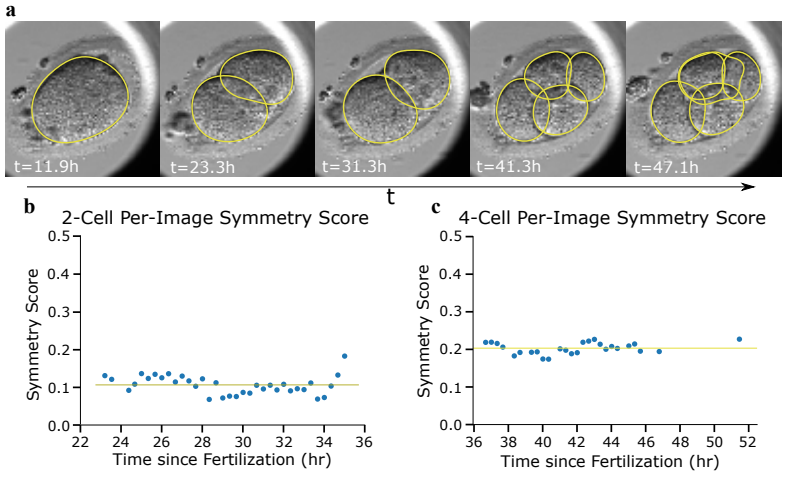

Figure A.2: Cell symmetry measurements. (**A**) Time-lapse video of a cleavage stage embryo with cell contour detection highlighted in yellow. Per-image (**B**) 2-cell and (**C**) 4-cell stage symmetry score measurements of the embryo shown in (**A**) (in blue dots) over time with the average symmetry score highlighted in yellow.

Supplementary Table S3. Patient demographics and cycle information.

	dataset (32, 939 embryos, 3, 921 cycles)
Patient age (*years*)	37.0 ± 5.6
BMI (kg/m^2)	24.4 ± 5.2
Oocytes retrieved (n)	10.2 ± 6.7
Fertilization method[a]	
ICSI	62.8%
IVF	37.2%
MII oocytes[b] (n)	8.8 ± 6.2
Normally fertilized[c] embryos (n)	
ICSI	6.5 ± 5.1
IVF	4.8 ± 3.9
Embryo transferred[d] (n)	1.3 ± 1.2
Fetal heartbeat[e] (%)	17.4
Live birth[f] (%)	15.8

Table A.3: Data are presented in mean ± SD or percentage. ICSI: intracytoplasmic sperm injection; MII: metaphase II oocyte. [a]Calculated per oocyte. [b]ICSI cycles only. [c]2PN. [d]Calculated per cycle. [e]Percentage of cycles with at least 1 embryo transferred that resulted in fetal heartbeat. [f]Percentage of cycles with at least 1 embryo transferred that resulted in live birth.

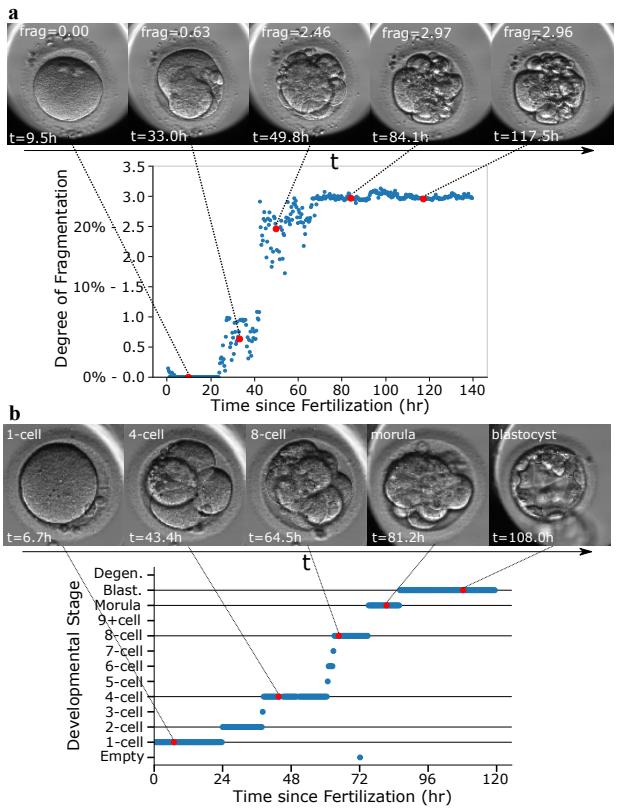

Figure A.3: Fragmentation and developmental stage detection results. (**A**) Time-lapse video and the network detection results of a fragmented embryo. The network output of degree of fragmentation for each image is highlighted in each image (top). Network detection of degree of fragmentation in blue dots of the fragmented embryo over time, with red dots corresponding to individual images the time lapse video (bottom). (**B**) Time-lapse video and network detection results of the developmental stage classifier. The network output of developmental stage for each image is highlighted in white (top). Network prediction of developmental stage in blue dots of the embryo over time, with red dots corresponding to individual images the time lapse video (bottom).

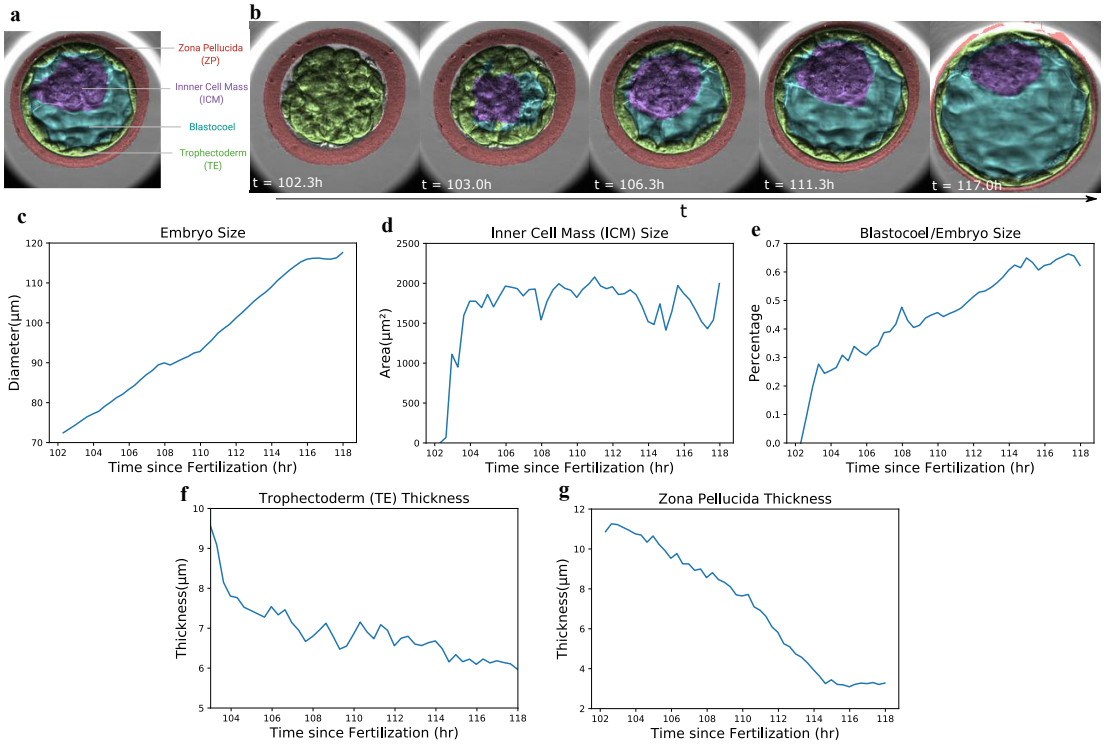

Figure A.4: Segmentation results and five measurements on an blastocyst stage embryo. (**A**) Schematic of blastocyst segmentation, with zona pellucida (ZP) highlighted in red, inner cell mass (ICM) in purple, blastocoel in cyan, and trophectoderm (TE) in green. (**B**) Time-lapse video of a blastocyst stage embryo with network detection of segmentation highlighted. Measurements of blastocyst size in terms of embryo size (not including ZP) (**C**), ICM size (**D**), and percentage of blastocoel area (**E**) of the same embryo over the entire blastocyst stage. Measurement of TE thickness (**F**) and ZP thickness (**G**).

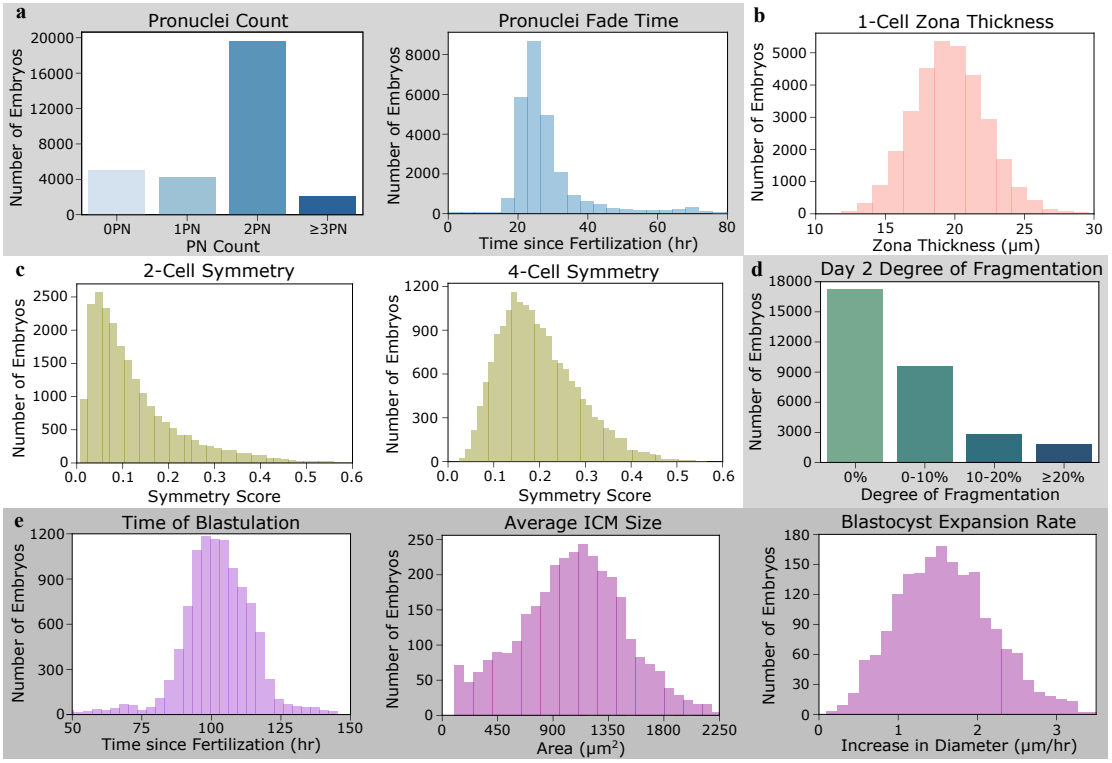

Figure A.5: Distributions of BlastAssist outputs of clinical features on all applicable data in the entire clinical dataset. (**A**) Pronuclei (PN) count (n = 31 035) (left) and PN fade time (n = 28 531) (right) measurements on embryos. (**B**) 1-cell stage zona pellucida (ZP) thickness measurements on embryos (n = 31 960). (**C**) 2-cell (n = 21 842) (left) and 4-cell (n = 17 544) (right) symmetry measurements on embryos. (**D**) Day 2 degree of fragmentation measurements on embryos (n =31 552). (**E**) Time of blastulation (n = 10 406) (left), average inner cell mass (ICM) size (n = 3586) (center), and blastocyst expansion rate in terms of increase in blastocyst diameter in $\mu m/hr$ (n = 1807) (right) measurements on embryos.

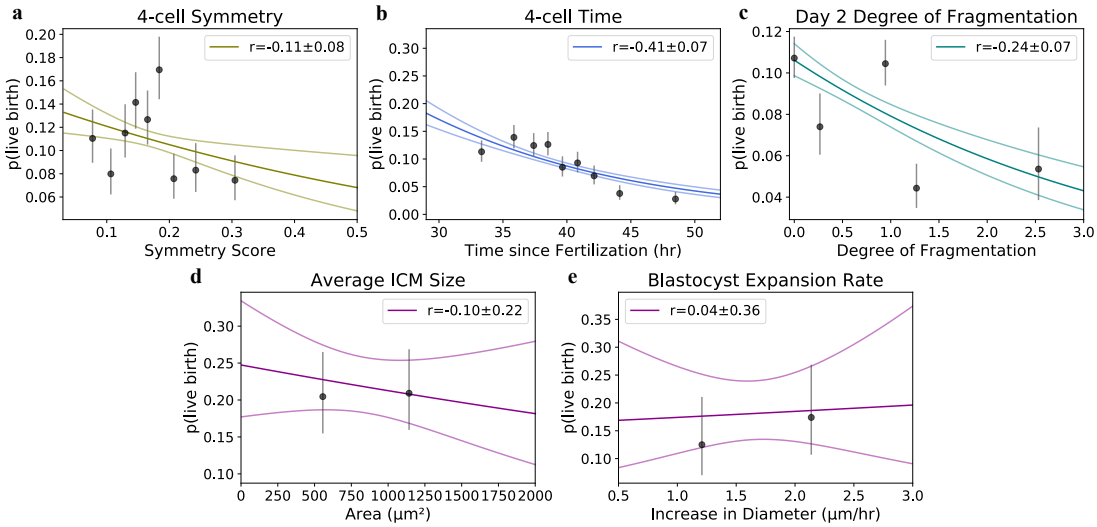

Figure A.6: Associations of additional BlastAssist outputs to clinical outcomes. Estimated probability of an embryo resulting in a live birth as a function of (**A**) 4-cell symmetry score ($n = 2338, t = -1.38, p < 0.2$). (**B**) time for embryos to reach 4-cell stage ($n = 3243, t = -5.86, p < 5 \times 10^{8}$). (**C**) degree of fragmentation on day 2 ($n = 3421, t = -6.71, p < 5 \times 10^{-11}$). (**D**) average inner cell mass (ICM) size during blastocyst stage ($n = 117$, $t = -0.45, p < 0.7$). (**E**) blastocyst expansion rate ($n = 48, t = 0.11, p < 0.95$). The data points and error bars show the probability of the embryo producing fetal heartbeat 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 $p(livebirth)$ and the variables are plotted in each figure. (See Supplementary Data File S1: Correlations to live birth.)

B

Supplementary Data of Chapter 3

B.1 Data Description

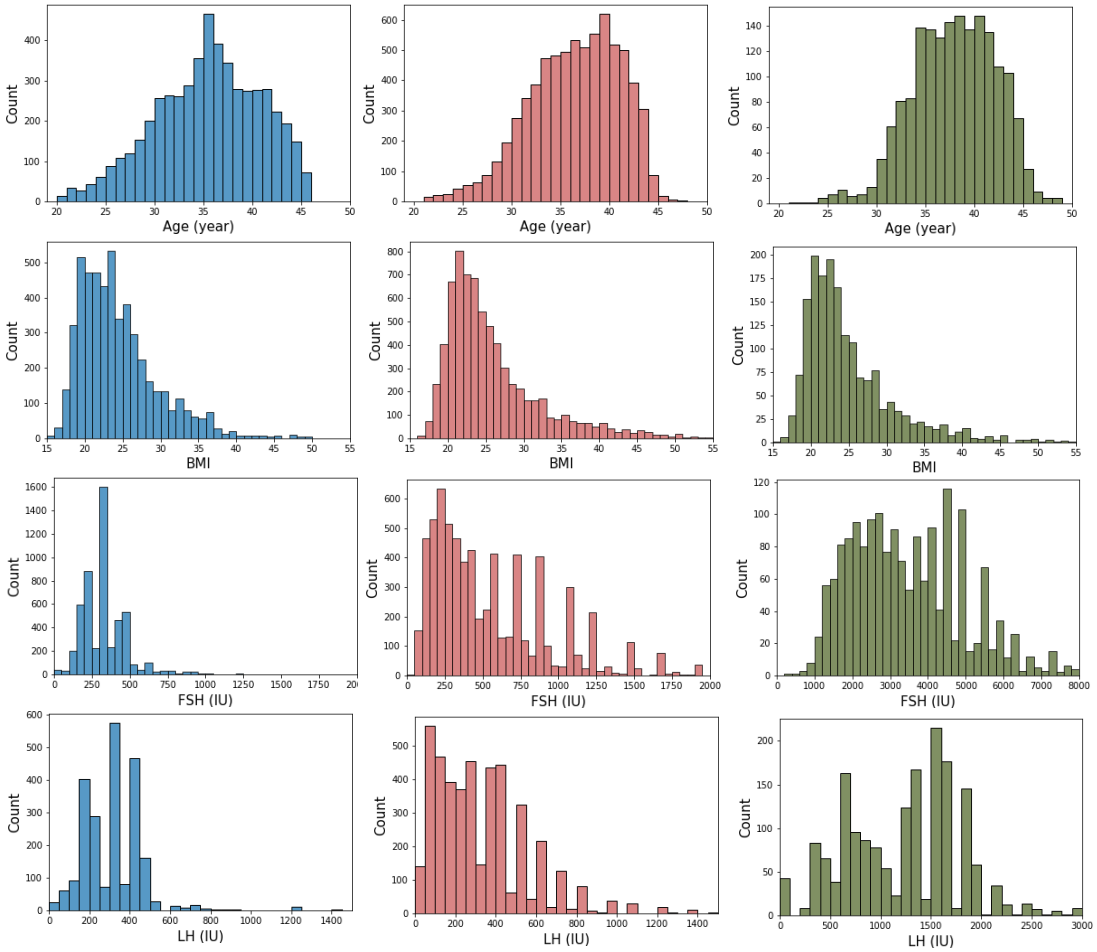

Figure B.1: Distribution of ovarian stimulation variables: Age, BMI, FSH, LH of the EHR dataset of Lis Maternity Hospital (LMH) (blue), Brigham and Women's Hospital (BWH) (red), and Weill Cornell Medicine (WCM) (green).

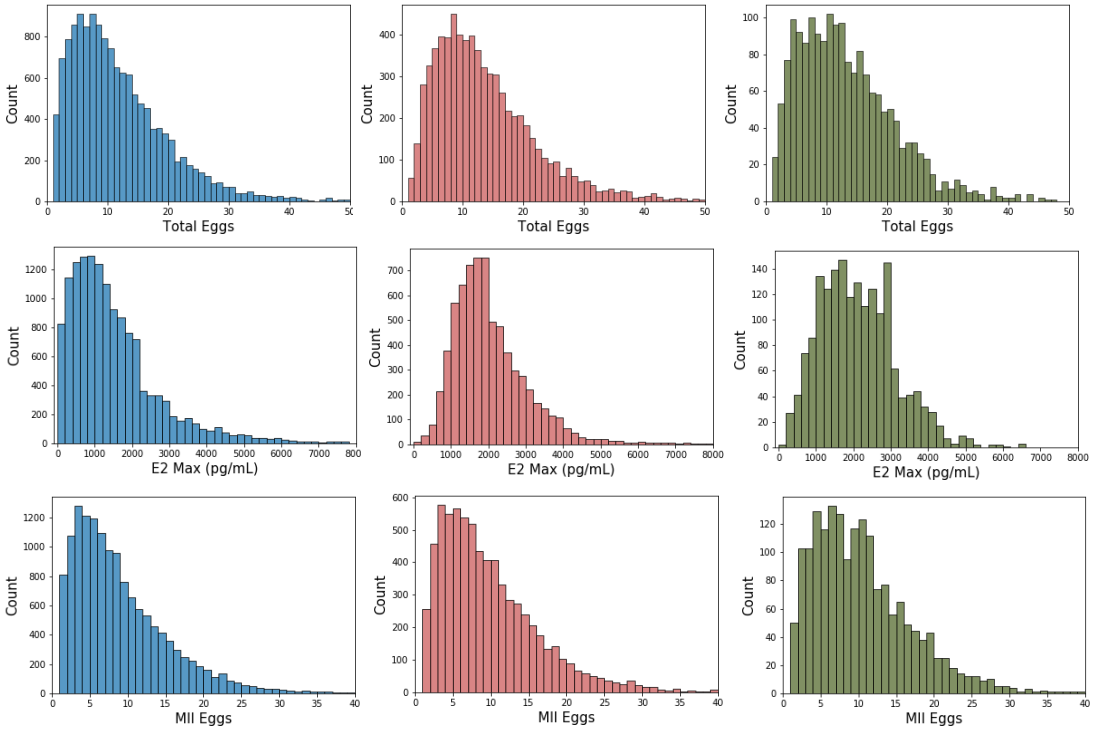

Figure B.2: Distribution of ovarian stimulation variables: Eggs, E2, MII of the EHR dataset of Lis Maternity Hospital (LMH) (blue), Brigham and Women's Hospital (BWH) (red), and Weill Cornell Medicine (WCM) (green).

Table B.1. Data distribution of discrete variables for fertilization model.

	Embryos (n)	
fert method	ICSI (126,383)	IVF (60,792)
mature status	MII (127,494)	non-MII (25,804)
pn count	2pn (120,181)	non-2pn (47,602)

Table B.1: dataset size in n embryos. ICSI: intracytoplasmic sperm injection; fert: fertilization; MII: metaphase II; pn: pronuclei.

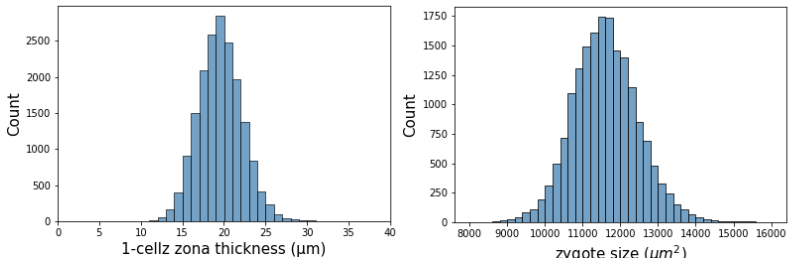

Figure B.3: Distribution of continuous variables for fertilization: 1-cell zona thickness, zygote size of dataset of Lis Maternity Hospital (LMH).

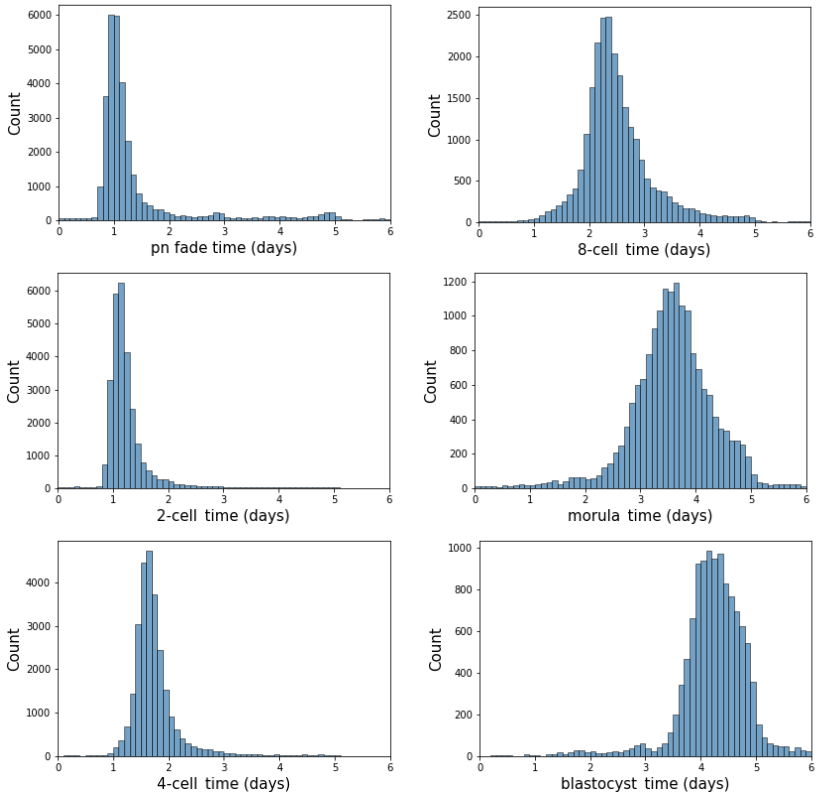

Figure B.4: Distribution of variables for fertilization: 2-cell time, 4-cell time, 8-cell time, morula time, blast time of dataset of Lis Maternity Hospital (LMH).

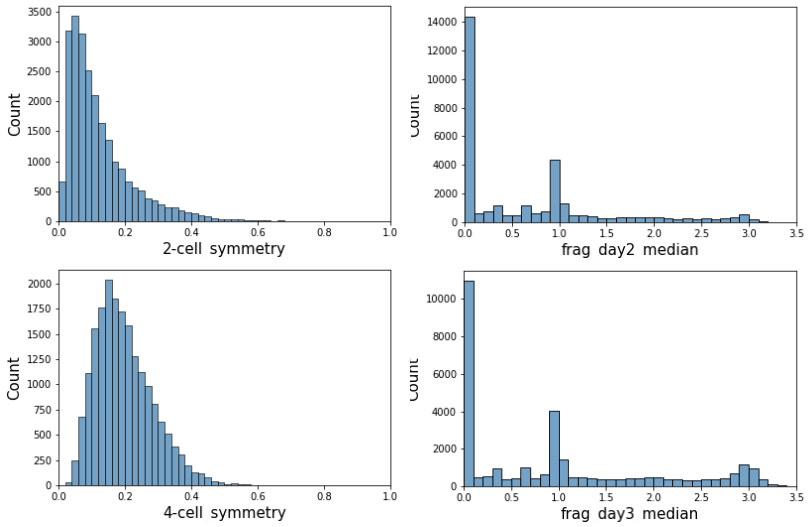

Figure B.5: Distribution of continuous variables for embryo development: 2-cell symm, 4-cell symm, d2 frag, d3 frag of dataset of Lis Maternity Hospital (LMH).

References

[1] Adamson, G., Zegers-Hochschild, F., Dyer, S., Chambers, G., de Mouzon, J., Ishihara, O., Kupka, M., Banker, M., Jwa, S., Elgindy, E., & Baker, V. (2022). International committee for monitoring assisted reproductive technology: World report on assisted reproductive technology, 2018.

[2] Afnan, M. A. M., Liu, Y., Conitzer, V., Rudin, C., Mishra, A., Savulescu, J., & Afnan, M. (2021a). Interpretable, not black-box, artificial intelligence should be used for embryo selection.

[3] Afnan, M. A. M., Rudin, C., Conitzer, V., Savulescu, J., Mishra, A., Liu, Y., & Afnan, M. (2021b). Ethical implementation of artificial intelligence to select embryos in in vitro fertilization. *arXiv preprint arXiv:2105.00060*.

[4] Ahlström, A., Lundin, K., Lind, A.-K., Gunnarsson, K., Westlander, G., Park, H., Thurin-Kjellberg, A., Thorsteinsdottir, S. A., Einarsson, S., Åström, M., et al. (2022). A double-blind randomized controlled trial investigating a time-lapse algorithm for selecting day 5 blastocysts for transfer. *Human reproduction (Oxford, England)*, (pp. deac020).

[5] Alikani, M., Cohen, J., Tomkin, G., Garrisi, G. J., Mack, C., & Scott, R. T. (1999). Human embryo fragmentation in vitro and its implications for pregnancy and implantation. *Fertility and sterility*, 71(5), 836–842.

[6] Alper, M. M. & Fauser, B. C. (2017). Ovarian stimulation protocols for ivf: is more better than less? *Reproductive biomedicine online*, 34(4), 345–353.

[7] Amir, H., Barbash-Hazan, S., Kalma, Y., Frumkin, T., Malcov, M., Samara, N., Hasson, J., Reches, A., Azem, F., & Ben-Yosef, D. (2019). Time-lapse imaging reveals delayed development of embryos carrying unbalanced chromosomal translocations. *Journal of assisted reproduction and genetics*, 36(2), 315–324.

[8] Anifandis, G., Sutovsky, P., Turek, P. J., Chavez, S. L., Kunej, T., Messini, C. I., Schon, S. B., Mavroforou, A., Adashi, E. Y., & Krawetz, S. A. (2022). Bioethics in human embryology: the double-edged sword of embryo research. *Systems biology in reproductive medicine*, 68(3), 169–179.

[9] Armstrong, S., Bhide, P., Jordan, V., Pacey, A., Marjoribanks, J., & Farquhar, C. (2019). Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database of Systematic Reviews*.

[10] Barberet, J., Bruno, C., Valot, E., Antunes-Nunes, C., Jonval, L., Chammas, J., Choux, C., Ginod, P., Sagot, P., Soudry-Faure, A., et al. (2019). Can novel early non-invasive biomarkers of embryo quality be identified with time-lapse imaging to predict live birth? *Human Reproduction*, 34(8), 1439–1449.

[11] Bellman, R. (1966). Dynamic programming. *Science*, 153(3731), 34–37.

[12] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 37–40). Springer.

[13] Bergh, C. (2005). Single embryo transfer: a mini-review. *Human reproduction*, 20(2), 323–327.

[14] Bertrand, E., Van den Bergh, M., & Englert, Y. (1995). Fertilization and early embryology: does zona pellucida thickness influence the fertilization rate? *Human reproduction*, 10(5), 1189–1193.

[15] Bertrand, E., Van den Bergh, M., & Englert, Y. (1996). Clinical parameters influencing human zona pellucida thickness. *Fertility and sterility*, 66(3), 408–411.

[16] Bormann, C. L., Kanakasabapathy, M. K., Thirumalaraju, P., Gupta, R., Pooniwala, R., Kandula, H., Hariton, E., Souter, I., Dimitriadis, I., Ramirez, L. B., et al. (2020). Performance of a deep learning based neural network in the selection of human blastocysts for implantation. *Elife*, 9, e55301.

[17] Breu, H., Gil, J., Kirkpatrick, D., & Werman, M. (1995). Linear time euclidean distance transform algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5), 529–533.

[18] Broughton, D. E. & Moley, K. H. (2017). Obesity and female infertility: potential mediators of obesity's impact. *Fertility and sterility*, 107(4), 840–847.

[19] Campbell, A., Fishel, S., Bowman, N., Duffy, S., Sedler, M., & Hickman, C. F. L. (2013). Modelling a risk classification of aneuploidy in human embryos using non-invasive morphokinetics. *Reproductive biomedicine online*, 26(5), 477–485.

[20] Campbell, A., Fishel, S., & Laegdsmand, M. (2014). Aneuploidy is a key causal factor of delays in blastulation: author response to 'a cautionary note against aneuploidy risk assessment using time-lapse imaging'. *Reproductive biomedicine online*, 28(3), 279–283.

[21] CDC (2021). 2019 national summary and clinic table dataset.

[22] Chandra, A., Copen, C. E., & Stephen, E. H. (2013). *Infertility and impaired fecundity in the United States, 1982-2010: data from the National Survey of Family Growth*. Number 2013. US Department of Health and Human Services, Centers for Disease Control.

[23] Chapuis, A., Gala, A., Ferrières-Hoa, A., Mullet, T., Bringer-Deutsch, S., Vintejoux, E., Torre, A., & Hamamah, S. (2017). Sperm quality and paternal age: effect on blastocyst formation and pregnancy rates. *Basic and clinical andrology*, 27, 1–9.

[24] Chavez-Badiola, A., Farias, A. F.-S., Mendizabal-Ruiz, G., Garcia-Sanchez, R., Drakeley, A. J., & Garcia-Sandoval, J. P. (2020). Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. *Scientific reports*, 10(1), 1–6.

[25] Choucair, F., Younis, N., & Hourani, A. (2021). The value of the modern embryologist to a successful ivf system: revisiting an age-old question. *Middle East Fertility Society Journal*, 26(1), 15.

[26] Cohen, J., Alikani, M., Trowbridge, J., & Rosenwaks, Z. (1992). Implantation enhancement by selective assisted hatching using zona drilling of human embryos with poor prognosis. *Human Reproduction*, 7(5), 685–691.

[27] Coticchio, G., Mignini Renzini, M., Novara, P., Lain, M., De Ponti, E., Turchi, D., Fadini, R., & Dal Canto, M. (2018). Focused time-lapse analysis reveals novel aspects of human fertilization and suggests new parameters of embryo viability. *Human reproduction*, 33(1), 23–31.

[28] Cui, W. (2010). Mother or nothing: the agony of infertility.

[29] Cutting, R. (2018). Single embryo transfer for all. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 53, 30–37.

[30] Dai, J., He, K., & Sun, J. (2016). Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3150–3158).

[31] De Neubourg, D. & Gerris, J. (2003). Single embryo transfer–state of the art. *Reproductive biomedicine online*, 7(6), 615–622.

[32] Della Ragione, T., Verheyen, G., Papanikolaou, E. G., Van Landuyt, L., Devroey, P., & Van Steirteghem, A. (2007). 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. *Reproductive Biology and Endocrinology*, 5(1), 1–8.

[33] Diwekar, U., Patel, N., Patel, N., Bhandarka, H., Patel, M., Ghoghari, P., Vyas, K., & Joag, S. (2023). Ivf stimulation-personalized, optimized, and simplified using an advanced decision-support tool: A randomized trial. *Journal of IVF-Worldwide*, 1(1-3), 1–12.

[34] Dohrmann, C., Busby, H., & Trujillo, D. (1988). Smoothing noisy data using dynamic programming and generalized cross-validation. *Journal of Biomechanical Engineering*, 110(1), 37–41.

[35] Dolinko, A. V., Farland, L., Kaser, D., Missmer, S., & Racowsky, C. (2017). National survey on use of time-lapse imaging systems in ivf laboratories. *Journal of assisted reproduction and genetics*, 34(9), 1167–1172.

[36] Durán, J. M. & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. *Journal of Medical Ethics*, 47(5), 329–335.

[37] Ebner, T., Moser, M., Sommergruber, M., & Tews, G. (2003). Selection based on morphological assessment of oocytes and embryos at different stages of preimplantation development: a review. *Human reproduction update*, 9(3), 251–262.

[38] Elder, K. & Dale, B. (2020). *In-vitro fertilization*. Cambridge University Press.

[39] Filho, E. S., Noble, J., Poli, M., Griffiths, T., Emerson, G., & Wells, D. (2012). A method for semi-automatic grading of human blastocyst microscope images. *Human Reproduction*, 27(9), 2641–2648.

[40] Franasiak, J. M., Forman, E. J., Hong, K. H., Werner, M. D., Upham, K. M., Treff, N. R., & Scott Jr, R. T. (2014). The nature of aneuploidy with increasing age of the female partner: a review of 15,169 consecutive trophectoderm biopsies evaluated with comprehensive chromosomal screening. *Fertility and sterility*, 101(3), 656–663.

[41] Freedman, D., Pisani, R., & Purves, R. (2007). Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

[42] Gabrielsen, A., Bhatnager, P. R., Petersen, K., & Lindenberg, S. (2000). Influence of zona pellucida thickness of human embryos on clinical pregnancy outcome following in vitro fertilization treatment. *Journal of assisted reproduction and genetics*, 17, 323–328.

[43] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

[44] Hagemann, A. R., Lanzendorf, S. E., Jungheim, E. S., Chang, A. S., Ratts, V. S., & Odem, R. R. (2010). A prospective, randomized, double-blinded study of assisted hatching in women younger than 38 years undergoing in vitro fertilization. *Fertility and sterility*, 93(2), 586–591.

[45] Hardy, M. & Reynolds, J. (2004). Incorporating categorical information into regression models: The utility of dummy variables. *Handbook of data analysis*, (pp. 229–55).

[46] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).

[47] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

[48] Hoffman, R. & Gross, L. (1975). Modulation contrast microscope. *Applied Optics*, 14(5), 1169–1176.

[49] Homer, H. A. (2019). Preimplantation genetic testing for aneuploidy (pgt-a): The biology, the technology and the clinical outcomes. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, 59(2), 317–324.

[50] Jang, W.-D., Wei, D., Zhang, X., Leahy, B., Yang, H., Tompkin, J., Ben-Yosef, D., Needleman, D., & Pfister, H. (2020). Learning vector quantized shape code for amodal blastomere instance segmentation. *arXiv preprint arXiv:2012.00985*.

[51] Jordan, M. I. & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260.

[52] Khan, A., Gould, S., & Salzmann, M. (2016). Deep convolutional neural networks for human embryonic cell counting. In *European Conference on Computer Vision* (pp. 339–348).: Springer.

[53] Kheradmand, S., Singh, A., Saeedi, P., Au, J., & Havelock, J. (2017). Inner cell mass segmentation in human hmc embryo images using fully convolutional network. In *2017 IEEE International Conference on Image Processing (ICIP)* (pp. 1752–1756).: IEEE.

[54] Khosravi, P., Kazemi, E., Zhan, Q., Malmsten, J. E., Toschi, M., Zisimopoulos, P., Sigaras, A., Lavery, S., Cooper, L. A., Hickman, C., et al. (2019). Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ digital medicine*, 2(1), 1–9.

[55] Kirkegaard, K., Agerholm, I. E., & Ingerslev, H. J. (2012). Time-lapse monitoring as a tool for clinical embryo assessment. *Human reproduction*, 27(5), 1277–1285.

[56] Knoppers, B. M., Bordet, S., & Isasi, R. (2009). The human embryo: ethical and legal aspects. *Human embryogenesis: methods and protocols*, (pp. 281–305).

[57] Kobayashi, T., Ishikawa, H., Ishii, K., Sato, A., Nakamura, N., Saito, Y., Hasegawa, H., Fujita, M., Mitsuhashi, A., & Shozu, M. (2021). Time-lapse monitoring of fertilized human oocytes focused on the incidence of opn embryos in conventional in vitro fertilization cycles. *Scientific Reports*, 11(1), 18862.

[58] Koifman, M., Lahav-Baratz, S., Shopen, L., Idit, B., Ishai, D., Wiener-Megnazi, Z., Auslender, R., Dirnfeld, M., et al. (2014). In vitro fertilization outcomes following assisted hatching of embryos with thick zona pellucida—a prospective randomized study. *Advances in Reproductive Sciences*, 2(04), 76.

[59] Koller, D. & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

[60] Kovacs, P. & Lieman, H. J. (2019). Which embryo selection method should be offered to the patients? *Journal of Assisted Reproduction and Genetics*, 36(4), 603–605.

[61] Kragh, M. F., Rimestad, J., Berntsen, J., & Karstoft, H. (2019). Automatic grading of human blastocysts from time-lapse imaging. *Computers in biology and medicine*, 115, 103494.

[62] Lau, T., Ng, N., Gingold, J., Desai, N., McAuley, J., & Lipton, Z. C. (2019). Embryo staging with weakly-supervised region selection and dynamically-decoded predictions. *arXiv preprint arXiv:1904.04419*.

[63] Leahy, B. D., Jang, W.-D., Yang, H. Y., Struyven, R., Wei, D., Sun, Z., Lee, K. R., Royston, C., Cam, L., Kalma, Y., et al. (2020). 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, October 4–8, 2020, Proceedings, Part V 23* (pp. 25–35).: Springer.

[64] Leahy, B. D., Racowsky, C., & Needleman, D. (2021). Inferring simple but precise quantitative models of human oocyte and early embryo development. *Interface*, 18(20210475).

[65] Lee, A. M., Connell, M. T., Csokmay, J. M., & Styer, A. K. (2016). Elective single embryo transfer-the power of one. *Contraception and reproductive medicine*, 1(1), 11.

[66] Lee, E., Illingworth, P., Wilton, L., & Chambers, G. M. (2015). The clinical effectiveness of preimplantation genetic diagnosis for aneuploidy in all 24 chromosomes (pgd-a): systematic review. *Human Reproduction*, 30(2), 473–483.

[67] Lee, M.-J., Lee, R. K.-K., Lin, M.-H., & Hwu, Y.-M. (2012). Cleavage speed and implantation potential of early-cleavage embryos in ivf or icsi cycles. *Journal of assisted reproduction and genetics*, 29(8), 745–750.

[68] Lemmen, J., Agerholm, I., & Ziebe, S. (2008). Kinetic markers of human embryo quality using time-lapse recordings of ivf/icsi-fertilized oocytes. *Reproductive biomedicine online*, 17(3), 385–391.

[69] Lewis, E. I., Farhadifar, R., Farland, L. V., Needleman, D. J., Missmer, S. A., & Racowsky, C. (2017). Use of imaging software for assessment of the associations among zona pellucida thickness variation, assisted hatching, and implantation of day 3 embryos. *Journal of assisted reproduction and genetics*, 34(10), 1261–1269.

[70] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).: Springer.

[71] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

[72] Lukyanenko, S., Jang, W.-D., Wei, D., Struyven, R., Kim, Y., Leahy, B., Yang, H., Rush, A., Ben-Yosef, D., Needleman, D., et al. (2021). Developmental stage classification of embryos using two-stream neural network with linear-chain conditional random field. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 363–372).: Springer.

[73] Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Büttner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, A. T., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Benis, A., Blaschko, M., Cardoso, M. J., Cheplygina, V., Cimini, B. A., Collins, G. S., Farahani, K., Ferrer, L., Galdran, A., van Ginneken, B., Haase, R., Hashimoto, D. A., Hoffman, M. M., Huisman, M., Jannin, P., Kahn, C. E., Kainmueller, D., Kainz, B., Karargyris, A., Karthikesalingam, A., Kenngott, H., Kofler, F., Kopp-Schneider, A., Kreshuk, A., Kurc, T., Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Mattson, P., Meijering, E., Menze, B., Moons, K. G. M., Müller, H., Nichyporuk, B., Nickel, F., Petersen, J., Rajpoot, N., Rieke, N., Saez-Rodriguez, J., Sánchez, C. I., Shetty, S., van Smeden, M., Summers, R. M., Taha, A. A., Tiulpin, A., Tsaftaris, S. A., Calster, B. V., Varoquaux, G., & Jäger, P. F. (2023). Metrics reloaded: Recommendations for image analysis validation.

[74] Malmsten, J., Zaninovic, N., Zhan, Q., Rosenwaks, Z., & Shan, J. (2019). 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)* (pp. 1–4).: IEEE.

[75] Marco-Jiménez, F., Naturil-Alfonso, C., Jiménez-Trigos, E., Lavara, R., & Vicente, J. (2012). Influence of zona pellucida thickness on fertilization, embryo implantation and birth. *Animal reproduction science*, 132(1-2), 96–100.

[76] Mastenbroek, S., Scriven, P., Twisk, M., Viville, S., Van der Veen, F., & Repping, S. (2008). What next for preimplantation genetic screening? more randomized controlled trials needed? *Human reproduction*, 23(12), 2626–2628.

[77] Mastenbroek, S., van der Veen, F., Aflatoonian, A., Shapiro, B., Bossuyt, P., & Repping, S. (2011). Embryo selection in ivf. *Human Reproduction*, 26(5), 964–966.

[78] McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2), 143–149.

[79] Morales, D. A., Bengoetxea, E., & Larranaga, P. (2008). Automatic segmentation of zona pellucida in human embryo images applying an active contour model. *Proc. of MIUA*.

[80] Nickkho-Amiry, M., Horne, G., Akhtar, M., Mathur, R., & Brison, D. (2019). Hydatidiform molar pregnancy following assisted reproduction. *Journal of assisted reproduction and genetics*, 36(4), 667–671.

[81] Norwitz, E. R., Edusa, V., & Park, J. S. (2005). Maternal physiology and complications of multiple pregnancy. *Seminars in perinatology*, 29(5), 338–348.

[82] of the American Society for Reproductive Medicine, P. C. (2017). Guidance on the limits to the number of embryos to transfer: a committee opinion. *Fertility and sterility*, 107(4), 901.

[83] Paternot, G., Wetzels, A. M., Thonon, F., Vansteenbrugge, A., Willemen, D., Devroe, J., Debrock, S., D'Hooghe, T. M., & Spiessens, C. (2011). Intra-and interobserver analysis in the morphological assessment of early stage embryos during an ivf procedure: a multicentre study. *Reproductive biology and endocrinology*, 9(1), 1–5.

[84] Paulson, R. J. (2020). Hidden in plain sight: the overstated benefits and underestimated losses of potential implantations associated with advertised pgt-a success rates. *Human Reproduction*, 35(3), 490–493.

[85] Petersen, B. M., Boel, M., Montag, M., & Gardner, D. K. (2016). Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on day 3. *Human reproduction*, 31(10), 2231–2244.

[86] Poon, A. I. & Sung, J. J. (2021). Opening the black box of ai-medicine. *Journal of gastroenterology and hepatology*, 36(3), 581–584.

[87] Popovic, M., Dhaenens, L., Boel, A., Menten, B., & Heindryckx, B. (2020). Chromosomal mosaicism in human blastocysts: the ultimate diagnostic dilemma. *Human Reproduction Update*, 26(3), 313–334.

[88] Racowsky, C., Stern, J. E., Gibbons, W. E., Behr, B., Pomeroy, K. O., & Biggers, J. D. (2011). 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. *Fertility and sterility*, 95(6), 1985–1989.

[89] Rad, R. M., Saeedi, P., Au, J., & Havelock, J. (2018a). A hybrid approach for multiple blastomeres identification in early human embryo images. *Computers in biology and medicine*, 101, 100–111.

[90] Rad, R. M., Saeedi, P., Au, J., & Havelock, J. (2018b). 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)* (pp. 3518–3522).: IEEE.

[91] Rad, R. M., Saeedi, P., Au, J., & Havelock, J. (2020). Trophectoderm segmentation in human embryo images via inceptioned u-net. *Medical image analysis*, 62, 101612.

[92] Raju, G. A. R., Chavan, R., Deenadayal, M., Gunasheela, D., Gutgutia, R., Haripriya, G., Govindarajan, M., Patel, N. H., & Patki, A. S. (2013). Luteinizing hormone and follicle stimulating hormone synergy: A review of role in controlled ovarian hyper-stimulation. *Journal of human reproductive sciences*, 6(4), 227–234.

[93] Rittenberg, V., Seshadri, S., Sunkara, S. K., Sobaleva, S., Oteng-Ntim, E., & El-Toukhy, T. (2011). Effect of body mass index on ivf treatment outcome: an updated systematic review and meta-analysis. *Reproductive biomedicine online*, 23(4), 421–439.

[94] Rosenwaks, Z., Davis, O. K., & Damario, M. A. (1995). The role of maternal age in assisted reproduction. *Human Reproduction*, 10(suppl_1), 165–173.

[95] Rubio, I., Galán, A., Larreategui, Z., Ayerdi, F., Bellver, J., Herrero, J., & Meseguer, M. (2014). Clinical validation of embryo culture and selection by morphokinetic analysis: a randomized, controlled trial of the embryoscope. *Fertility and sterility*, 102(5), 1287–1294.

[96] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

[97] Sciorio, R. & Dattilo, M. (2020). Pgt-a preimplantation genetic testing for aneuploidies and embryo selection in routine art cycles: Time to step back? *Clinical Genetics*, 98(2), 107–115.

[98] Segawa, T., Teramoto, S., Omi, K., Miyauchi, O., Watanabe, Y., & Osada, H. (2015). Changes in estrone and estradiol levels during follicle development: a retrospective large-scale study. *Reproductive Biology and Endocrinology*, 13, 1–7.

[99] Sela, R., Samuelov, L., Almog, B., Schwartz, T., Cohen, T., Amit, A., Azem, F., & Ben-Yosef, D. (2012). An embryo cleavage pattern based on the relative blastomere size as a function of cell number for predicting implantation outcome. *Fertility and sterility*, 98(3), 650–656.

[100] Sermondade, N., Huberlant, S., Bourhis-Lefebvre, V., Arbo, E., Gallot, V., Colombani, M., & Fréour, T. (2019). Female obesity is negatively associated with live birth rate following ivf: a systematic review and meta-analysis. *Human reproduction update*, 25(4), 439–451.

[101] Sfakianoudis, K., Maziotis, E., Grigoriadis, S., Pantou, A., Kokkini, G., Trypidi, A., Giannelou, P., Zikopoulos, A., Angeli, I., Vaxevanoglou, T., et al. (2022). Reporting on the value of artificial intelligence in predicting the optimal embryo for transfer: A systematic review including data synthesis. *Biomedicines*, 10(3), 697.

[102] Silver, D. H., Feder, M., Gold-Zamir, Y., Polsky, A. L., Rosentraub, S., Shachor, E., Weinberger, A., Mazur, P., Zukin, V. D., & Bronstein, A. M. (2020). Data-driven prediction of embryo implantation probability using ivf time-lapse imaging. *arXiv preprint arXiv:2006.01035*.

[103] Simopoulou, M., Sfakianoudis, K., Maziotis, E., Antoniou, N., Rapani, A., Anifandis, G., Bakas, P., Bolaris, S., Pantou, A., Pantos, K., et al. (2018). Are computational applications the "crystal ball" in the ivf laboratory? the evolution from mathematics to artificial intelligence. *Journal of Assisted Reproduction and Genetics*, 35(9), 1545–1557.

[104] Somigliana, E., Busnelli, A., Paffoni, A., Vigano, P., Riccaboni, A., Rubio, C., & Capalbo, A. (2019). Cost-effectiveness of preimplantation genetic testing for aneuploidies. *Fertility and Sterility*, 111(6), 1169–1176.

[105] Student (1908). The probable error of a mean. *Biometrika*, (pp. 1–25).

[106] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

[107] Tian, T., Kong, F., Yang, R., Long, X., Chen, L., Li, M., Li, Q., Hao, Y., He, Y., Zhang, Y., et al. (2023). A bayesian network model for prediction of low or failed fertilization in assisted reproductive technology based on a large clinical real-world data. *Reproductive Biology and Endocrinology*, 21(1), 8.

[108] Tran, D., Cooke, S., Illingworth, P., & Gardner, D. (2019). Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. *Human Reproduction*, 34(6), 1011–1018.

[109] Ubaldi, F. M., Cimadomo, D., Vaiarelli, A., Fabozzi, G., Venturella, R., & Mazzilli, R. (2019). Advanced maternal age in ivf: still a challenge? the present and the future of its treatment. *Frontiers in endocrinology*, 10, 406173.

[110] Uyar, A., Bener, A., Ciray, H. N., & Bahceci, M. (2010). Bayesian networks for predicting ivf blastocyst development. In *2010 20th International Conference on Pattern Recognition* (pp. 2772–2775).: IEEE.

[111] van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1), 1.

[112] Van Loendersloot, L., Van Wely, M., Limpens, J., Bossuyt, P., Repping, S., & Van Der Veen, F. (2010). Predictive factors in in vitro fertilization (ivf): a systematic review and meta-analysis. *Human reproduction update*, 16(6), 577–589.

[113] van Montfoort, A. P., Dumoulin, J. C., Land, J. A., Coonen, E., Derhaag, J. G., & Evers, J. L. (2005). Elective single embryo transfer (eset) policy in the first three ivf/icsi treatment cycles. *Human Reproduction*, 20(2), 433–436.

[114] VerMilyea, M., Hall, J., Diakiw, S., Johnston, A., Nguyen, T., Perugini, D., Miller, A., Picou, A., Murphy, A., & Perugini, M. (2020). Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during ivf. *Human Reproduction*, 35(4), 770–784.

[115] Viotti, M. (2020). Preimplantation genetic testing for chromosomal abnormalities: aneuploidy, mosaicism, and structural rearrangements. *Genes*, 11(6), 602.

[116] Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

[117] Weitzman, V. N., Schnee-Riesz, J., Benadiva, C., Nulsen, J., Siano, L., & Maier, D. (2010). Predictive value of embryo grading for embryos with known outcomes. *Fertility and sterility*, 93(2), 658–662.

[118] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492–1500).

[119] Yan, J., Wu, K., Tang, R., Ding, L., & Chen, Z.-J. (2012). Effect of maternal age on the outcomes of in vitro fertilization and embryo transfer (ivf-et). *Science China Life Sciences*, 55, 694–698.

[120] Yang, H. Y., Leahy, B. D., Jang, W.-D., Wei, D., Kalma, Y., Rahav, R., Carmon, A., Kopel, R., Azem, F., Venturas, M., et al. (2024). Blastassist: a deep learning pipeline to measure interpretable features of human embryos. *Human Reproduction*, (pp. deae024).

[121] Yousuf Harun, M., Huang, T., & Ohta, A. T. (2020). Inner cell mass and trophectoderm segmentation in human blastocyst images using deep neural network. *arXiv e-prints*, (pp. arXiv–2008).

[122] Zhang, X., Wang, Y., Zhao, N., Liu, P., & Huang, J. (2020). Variations in chromosomal aneuploidy rates in ivf blastocysts and early spontaneous abortion chorionic villi. *Journal of assisted reproduction and genetics*, 37, 527–537.

[123] Ziebe, S., Lundin, K., Loft, A., Bergh, C., Nyboe Andersen, A., Selleskog, U., Nielsen, D., Grøndahl, C., Kim, H., & Arce, J.-C. (2003). 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. *Human Reproduction*, 18(12), 2575–2581.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.