HHS Public Access

Author manuscript

Lancet Digit Health. Author manuscript; available in PMC 2023 May 18.

Published in final edited form as:

Lancet Digit Health. 2023 January ; 5(1): e28–e40. doi:10.1016/S2589-7500(22)00213-8.

A non-invasive artificial intelligence approach for the prediction of human blastocyst ploidy: a retrospective model development and validation study

Josue Barnes,

Matthew Brendel,

Vianne R Gao,

Suraj Rajendran,

Junbum Kim,

Qianzi Li,

Jonas E Malmsten,

Jose T Sierra,

Pantelis Zisimopoulos,

Alexandros Sigaras,

Pegah Khosravi,

Marcos Meseguer,

Qiansheng Zhan,

Zev Rosenwaks,

Olivier Elemento,

Nikica Zaninovic,

Iman Hajirasouliha

Department of Physiology and Biophysics (J Barnes BS, M Brendel MEng, S Rajendran BS, J Kim MEng, P Zisimopoulos MSc, A Sigaras MSc, P Khosravi PhD, Prof O Elemento PhD, I

Correspondence to: Dr Iman Hajirasouliha, Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA, imh2003@med.cornell.edu.

Contributors

JB, PK, NZ, OE, and IH conceived the study. JB, MB, VRG, and IH conceived the method and designed the algorithmic techniques. JB wrote the codes and did the computational analysis with input from IH QZ, and JEM, and NZ provided the primary and WCM-ES+ datasets and labelled images from Weill Cornell Medicine. MM provided the IVI Valencia dataset and labelled images. PZ and AS designed the user interface. MB, VRG, SR, JK, QL, JTS, and PK also contributed to computational analysis and validations. ZR provided critical reading and suggestions. JB drafted the manuscript with input from MB, QZ, OE, NZ, and IH. All the authors read the paper and suggested edits. IH supervised the project. IH, JB, and MB accessed and verified the data. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

All other authors declare no competing interests.

Data sharing

The embryo-imaging datasets analysed in this study are not publicly available, owing to reasonable privacy and security concerns. The embryo-imaging data are not easily redistributable to researchers other than those engaged in the institutional review board-approved research collaborations with the named medical centres. Our method is not specific to the datasets used in this study and users can train and test the deep-learning model on any relevant imaging data. The official source code repository is publicly available on Github (https://github.com/ihlab/stork-a). STORK-A is available through a web-based user interface (stork-a.eipm-research.org; to gain access to this password protected site for research purposes only, contact the corresponding author).

Hajirasouliha PhD) and Institute for Computational Biomedicine (J Barnes, M Brendel, V R Gao BS, S Rajendran, J Kim, Q Li BS, P Zisimopoulos, A Sigaras, P Khosravi, Prof O Elemento, I Hajirasouliha) and Ronald O Perelman and Claudia Cohen Center for Reproductive Medicine (J E Malmsten DPS, Q Zhan PhD, Prof Z Rosenwaks MD, N Zaninovic PhD) and Caryl and Israel Englander Institute for Precision Medicine (P Zisimopoulos, A Sigaras, P Khosravi, Prof O Elemento, I Hajirasouliha) and Meyer Cancer Center (Prof O Elemento, I Hajirasouliha) and WorldQuant Initiative for Quantitative Prediction (Prof O Elemento), Weill Cornell Medicine, New York, NY, USA; Tri-Institutional Computational Biology & Medicine Program, Cornell University, NY, USA (V R Gao, S Rajendran, Q Li); QED Analytics, Princeton, NJ, USA (J T Sierra PhD); Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA (P Khosravi); IVI Valencia, Health Research Institute la Fe, Valencia, Spain (M Meseguer PhD)

Summary

Background—One challenge in the field of in-vitro fertilisation is the selection of the most viable embryos for transfer. Morphological quality assessment and morphokinetic analysis both have the disadvantage of intra-observer and inter-observer variability. A third method, preimplantation genetic testing for aneuploidy (PGT-A), has limitations too, including its invasiveness and cost. We hypothesised that differences in aneuploid and euploid embryos that allow for model-based classification are reflected in morphology, morphokinetics, and associated clinical information.

Methods—In this retrospective study, we used machine-learning and deep-learning approaches to develop STORK-A, a non-invasive and automated method of embryo evaluation that uses artificial intelligence to predict embryo ploidy status. Our method used a dataset of 10 378 embryos that consisted of static images captured at 110 h after intracytoplasmic sperm injection, morphokinetic parameters, blastocyst morphological assessments, maternal age, and ploidy status. Independent and external datasets, Weill Cornell Medicine EmbryoScope+ (WCM-ES+; Weill Cornell Medicine Center of Reproductive Medicine, NY, USA) and IVI Valencia (IVI Valencia, Health Research Institute la Fe, Valencia, Spain) were used to test the generalisability of STORK-A and were compared measuring accuracy and area under the receiver operating characteristic curve (AUC).

Findings—Analysis and model development included the use of 10 378 embryos, all with PGT-A results, from 1385 patients (maternal age range 21–48 years; mean age 36·98 years [SD 4·62]). STORK-A predicted aneuploid versus euploid embryos with an accuracy of 69·3% (95% CI 66·9–71·5; AUC 0·761; positive predictive value [PPV] 76·1%; negative predictive value [NPV] 62·1%) when using images, maternal age, morphokinetics, and blastocyst score. A second classification task trained to predict complex aneuploidy versus euploidy and single aneuploidy produced an accuracy of 74·0% (95% CI 71·7–76·1; AUC 0·760; PPV 54·9%; NPV 87·6%) using an image, maternal age, morphokinetic parameters, and blastocyst grade. A third classification task trained to predict complex aneuploidy versus euploidy had an accuracy of 77·6% (95% CI 75·0–80·0; AUC 0·847; PPV 76·7%; NPV 78·0%). STORK-A reported accuracies of 63·4% (AUC 0·702) on the WCM-ES+ dataset and 65·7% (AUC 0·715) on the IVI Valencia dataset, when using an image, maternal age, and morphokinetic parameters, similar to the STORK-A test dataset accuracy of 67·8% (AUC 0·737), showing generalisability.

Interpretation—As a proof of concept, STORK-A shows an ability to predict embryo ploidy in a non-invasive manner and shows future potential as a standardised supplementation to traditional methods of embryo selection and prioritisation for implantation or recommendation for PGT-A.

Funding—US National Institutes of Health.

Introduction

As women approach the end of their childbearing years the incidence of aneuploid embryos (ie, those that show chromosomal abnormalities), increases, which often results in serious clinical consequences such as infertility, miscarriage, and birth defects.[1] As a result, there is an increasing trend for couples to conceive using assisted reproductive technologies. Experts in the field of reproductive medicine have geared their efforts towards selecting and transferring the single most viable embryo that will result in the livebirth of one healthy child. Reductions in the number of embryos transferred confer several advantages for a patient including decreased overall health-care costs, minimising potential complications, and reducing the mental, physical, and emotional tax of repeated implantation failures and pregnancy losses. This selection process presents itself as one of the chief challenges in the field of in-vitro fertilisation (IVF).

Both non-invasive and invasive methods for embryo selection are currently implemented in fertility clinics. Embryonic morphology assessment by an expert embryologist at discrete timepoints on day 3 or day 5 of development has been the predominant non-invasive means of evaluating embryo quality and subsequent selection for transfer.[2,3] By focusing visual quality assessment on a set of morphological features that correlate with viability, embryologists assign embryo quality through a rubric-like grading that emphasises three aspects of blastocyst morphology: degree of blastocyst expansion and hatching status; the inner cell mass; and the trophectoderm grade. Time-lapse microscopy has gained traction as a supplemental tool for improved embryo selections. This technology allows embryologists to monitor embryo development and do morphokinetic analysis with increased ease, which have been shown to be associated with improved implantation potential and pregnancy rate.[4] Although morphological assessment and morphokinetic annotations are noninvasive, the two methods are time-consuming and can have the disadvantage of intra-observer and inter-observer variability, due to their inherent subjectivity.[5–9]

Advancements in comprehensive chromosome screening technologies such as preimplantation genetic testing for aneuploidy (PGT-A) provide a means of unbiased implantation potential by ensuring the transfer of a euploid, chromosomally normal embryo, which improves the chances of obtaining a successful livebirth. This method of embryo selection is especially useful for patients of advanced maternal age (eg, 35 years or older) who have an increased potentional for pregnancy failure (eg, miscarriage or stillbirth). Retrospective studies on PGT-A pregnancy outcomes using 620 cycles[10] and 1183 cycles,[11] and a randomised controlled trial of 661 women,[12] have shown that PGT-A increases implantation potential and pregnancy rates, particularly among patients of advanced maternal age.[10,12] However, evidence shows that among younger patients, the use of PGT-A did not show an improvement in these outcomes compared with not doing

PGT-A.[12] Although PGT-A can address issues of variability noted in morphological and morphokinetic analysis methods, several limitations remain. The invasive nature of PGT-A has given rise to moral and ethical issues and can result in a reduction in embryo quality and viability.[13–16] In addition to these limitations, doing PGT-A is costly and time-consuming, and requires a sophisticated molecular-genetics diagnostic laboratory and embryologists who have been expertly trained to minimise the number of blastocysts that undergo biopsy, reduce potential damage to the embryo to not reduce viability, and reduce embryo cryogenic damage.

The assessment of embryo quality using non-invasive imaging techniques requires a high degree of expertise and is subject to observer heterogeneity.[14–16] Deep-learning approaches have evolved over the past decade as a powerful tool for tasks such as image classification and are useful for the analysis of embryonic imaging data. Convolutional neural networks (CNNs) are one particular and widely used deep-learning approach for image classification tasks. Such networks are structured in different layers, and each layer consists of multiple image filters, which are used to extract important features, either from the raw image pixels in the first layer or from an intermediate representation of the image in subsequent layers. The filters are optimised to do the specific task of interest. In the past 4 years, studies have leveraged the use of artificial intelligence to automate the selection of embryos for IVF.

A study[17] of EmbryoScope time-lapse images from 98 blastocysts explored the use of morphokinetics to determine if aneuploids displayed substantial differences in temporal variables as euploids and subsequently modelled the risk of aneuploidy. The study identified that the time to the start of blastulation and the time to full blastulation were significantly different between euploid and aneuploid embryos.[17] With these results, the study authors created a simple recursive partitioning method to model the degree of aneuploid risk, resulting in an area under the receiver operating characteristic curve (AUC) of 0·72. However, several studies that sought to identify statistical differences in morphokinetics between euploid and aneuploid embryos, similar to the previous study,[17] resulted in conflicting outcomes with no single set of morphokinetic parameters consistently predicting embryo ploidy.[18–24] This absence of consistency can be because of inter-observer variability, the absence of standardised guidelines for annotating morphokinetic parameters, and embryo-culture protocols. Importantly, this absence places into question whether morphokinetics can reliably be used as an alternative to PGT-A for assessing ploidy status.

A study[25] using a total of 1231 embryo images examined the ability of their embryo-ranking intelligent classification algorithm (ERICA) deep-learning model to rank embryos using PGT-A results and β-HCG concentrations, thereby establishing good-prognosis and poor-prognosis ground-truth labels and maternal age. Ultimately, the study reported an accuracy of 70% (AUC 74%), a sensitivity of 54%, and a specificity of 86%. On a per-cycle basis, ERICA predicted a euploid embryo in the top rank in 15 (79%) of 19 cases, and a euploid embryo in the top two in 18 (95%) of 19 cycles. Within the study, the authors did not differentiate between single and complex aneuploids, assumed that β-HCG concentrations of 20 mUI/mL or higher on the seventh day of embryo development was euploid, and the dataset size was small for developing a robust and generalisable model.

A subsequent study[26] deviated from the common approach of using a single static image and instead used full-length time-lapse videos as proof of concept for ploidy prediction. Using a two-stream inflated three-dimensional ConvNet architecture with videos that spanned day 1–5 of development, the study achieved an AUC of 0·74 when predicting aneuploid versus euploid and mosaic embryos. The study's authors noted several limitations of their study, including dataset size (n=690), an absence of included embryos from patients older than 37 years, and an unbalanced dataset. An additional limitation, in terms of applicability and deployment in the clinic, was the use of time-lapse microscopy. Although there are certainly advantages to using time-lapse machinery, few clinics have it, which limits use of this method.

We propose a deep-learning method called STORK-A, which uses images captured by time-lapse microscopy and clinical information (eg, maternal age, morphokinetic parameters, and morphological assessment) to accurately predict human embryo ploidy. The purpose of STORK-A is to aid clinicians in the selection and prioritisation of embryos for PGT-A biopsy or implantation in a cost-efficient, standardised, and non-invasive manner.

Methods

Data source

In this retrospective study, machine-learning and deeplearning approaches were tested to develop a novel model for the prediction of ploidy status. The de-identified dataset consisted of static time-lapse images (500 × 500 pixels) at 110 h after intracytoplasmic sperm injection (ICSI), maternal age at the time of oocyte retrieval, blastocyst grade, blastocyst score, morphokinetic parameters ranging from pro-nuclear fading to the time of the start of blastulation, and PGT-A results. The dataset encompassed 10 378 human blastocysts (day 5 n=3994; day 6 n=6384) collected from 1385 patients from 2012 to 2017 at the Weill Cornell Medicine Center of Reproductive Medicine, New York, NY, USA.

Images and morphokinetics were captured using an EmbryoScope (Weill Cornell Medicine Center of Reproductive Medicine) time-lapse imaging instrument. Images taken at 110 h post ICSI were used because this was the average time at which embryologists assessed the morphological quality of embryos and biopsied cells for PGT-A. At 110 h, the developmental stage of embryos varies between the morula and blastocyst stage, thereby accounting for temporal differences in blastocyst formation. Four expertly trained embryologists at the Weill Cornell Medicine Center of Reproductive Medicine manually annotated morphokinetic parameters and assigned blastocyst grades using the Veeck and Zaninovic grading system,[27] which includes assessments of the inner cell mass, trophectoderm, and expansion. Blastocyst scores were derived from a system that converts trophectoderm, inner cell mass, and expansion grades into established numerical values.[28] The blasto cyst score also takes into consideration the day of blastocyst formation (day 5 *vs* day 6) when calculating the score.

Embryos were biopsied for PGT-A on day 5 if the morphological grade was 2BB or better using the Veeck and Zaninovic grading system,[27] which classifies embryos on the basis of the degree of blastocyst expansion (grades 1–6) and cell abundance and conformity in

both the trophectoderm (grades A–C) and the inner cell mass (grades A–C; an embryo that is 2BB or better is characterised as being in the blastocyst stage with a notable increase in size, thinning of the zona pellucida, trophectoderm with few but large cells, and a distinguishable inner cell mass with large loose cells). The remaining embryos were biopsied on day 6, as long as they had reached the blasto cyst stage. In instances in which patients had a small number of viable embryos, embryos were biopsied on day 6 if they were in the morula stage or the cavitating morula stage. PGT-A results were categorised into two classes, aneuploids (n=5953) and euploids (n=4425), and were used as ground truth labels for the ploidy prediction task. The aneuploid class could be further stratified into single aneuploids (n=2944) and complex aneuploids (n=3009). Single aneuploids have one chromosomal abnormality, whereas complex aneuploids have two or more. Livebirth outcomes and fetal heart results for 1638 transferred embryos were also included. Of the 10 378 embryos in the dataset, 2426 (single aneuploids n=697; complex aneuploids n=951; euploids n=778) had at least one or more morphokinetic parameters missing. For these instances of missing data, median imputation was used to replace missing morphokinetic parameters with median values. Additionally, 12 static images that were underexposed were removed from the dataset.

To confirm the generalisability of STORK-A, independent datasets were sourced: WCM-ES+ (from Weill Cornell Medicine, using the new EmbryoScope+ machines) and IVI Valencia (from IVI Valencia, Health Research Institute la Fe, Valencia, Spain). The WCM-ES+ dataset, captured from 2018 to 2019 using the EmbryoScope+, consisted of 841 embryos including single aneuploids (n=170), complex aneuploids (n=261), and euploids (n=410), maternal age, morphokinetic parameters, and morphological assessment. The IVI Valencia dataset from 2018 consisted of 554 embryos including aneuploids (n=319) and euploids (n=235), maternal age, and morphokinetic parameters. Morphokinetic parameters were manually annotated by embryologists at IVI Valencia. The morphological criteria to select embryos for biopsy were identical to those used by Weill Cornell Medicine.

Clinical feature importance determination using lasso regression

Lasso regression, with a regularisation term C of weight 0·01, was done using the scikit-learn package (version 1.1.1) in Python (version 3.7). The regularisation term was determined using a cross-validation grid search on each split of the training set with average precision as the scoring metric with regularisation terms starting at 0·0001 and increasing by a factor of 10 to a regularisation term of 10. Five-fold cross-validation was performed, and 95% CIs were calculated for three different tasks: aneuploids versus euploids; complex aneuploid versus euploids; and complex aneuploid versus single aneuploid plus euploids. Features included maternal age, morphokinetics from pro-nuclear fading to the time of the start of blastulation, blastocyst grade, which included assessment of trophectoderm, inner cell mass, the degree of expansion, and blastocyst score. All features were Z-score normalised on the basis of the training set for each of the cross-validation splits.

Morphological feature importance using logistic regression

Logistic regression was performed using the scikit-learn package with the following parameters: penalty=l2, C=10, and class_weight=balanced, where the penalty is the type

of regularisation, C is the strength of regularisation, and class_weight is the method for weighting the loss function to account for class imbalance. Two different sets of features were analysed on the basis of previous lasso regression results, maternal age, and blastocyst score, and each blastocyst grade component (trophectoderm, inner cell mass, and expansion). All grades were converted from letter to numerical grades, such that A=6, A=5, etc. Intermediate scores (ie, 1–2 for expansion grades), were given an intermediate score; in this case, 1·5. Z-score normalisation and median imput-ation were performed in the identical way as for the lasso regression. Three subsets of data based on ploidy type were compared: aneuploids versus euploids; complex aneuploid versus single aneuploid plus euploids; and complex aneuploid versus euploids. Five-fold cross-validation was performed, and validation set accuracy was reported (mean and 95% CIs). In addition, univariate analysis for each component of the blastocyst grades was done. The weights for each feature were recorded to analyse feature importance from the logistic regression models and the Shapley additive explanations values were also used to validate feature importance.[29]

Blastocyst score prediction

The embryologist-derived morphological assessments used in the study are subject to variability. To standardise morphological assessment, a blastocyst score regression model was trained using a deep-learning model, based on the ResNet18 architecture pretrained on ImageNet and performed using PyTorch (version 1.4.0). The ResNet18 architecture was modified to perform a regression task by adding two fully connected layers, both of which were fine-tuned to perform blastocyst score regression. Using the primary dataset, the model was trained and validated on images of embryos at 110 h after ICSI and used embryologist-derived blastocyst scores as ground truth labels to produce artificial intelligence blastocyst scores (AIBS). The model was trained for 20 epochs, with a batch size of 32, Adam optimisation with a learning rate of 0·001, and mean squared error loss. To ensure the model did not overfit on the training data, early stopping, with patience=2, was implemented.

Machine learning and deep learning

Extreme gradient boost decision tree (XGBoost), k-nearest neighbour (k-NN), support vector machine (SVM), and Random Forest were trained using five-fold cross-validation and tested in R (version 4.1.2) using the caret package (version 6.0–90). Clinical features were used for input with ploidy status determined by PGT-A as the predicted outcome.

To exploit the spatial features of static embryo images for ploidy classification, STORK-A was trained, validated, and tested using PyTorch (version 1.4.0). STORK-A is based on a ResNet18 CNN architecture pretrained on ImageNet.[30] The ResNet18 architecture was modified to concatenate features from images with clinical features before being passed on to two fully connected layers that were fine-tuned to output the predicted probabilities of a binary classification task (figure 1). Several models were created to assess combinations of feature input to identify which features performed best. Models were trained for 20 epochs, with a batch size of 32, Adam optimisation with a learning rate of 0·0001, and cross-entropy loss. Image augmentation was used to increase the magnitude of the training set and included a random resized crop of size 224 × 224 pixels and random horizontal and vertical flips.

A random 70:15:15 training, validation, and test (primary) split was applied and used consistently across all models by using a set seed for reproducibility and comparison across all models. Clinical features in the training, validation, and test sets of both machine-learning and deep-learning models were Z-score normalised to the training set. Several subsets of the data were identified to address different classification tasks, including: aneuploids (single aneuploid plus complex aneuploid) versus euploids; complex aneuploid versus single aneuploid plus euploids; and complex aneuploid versus euploids. To address issues of class imbalance in the training set, the minority class was oversampled.

All binary classification thresholds for both machine-learning and deep-learning models were maintained at 50%. For example, in the aneuploids versus euploids classification task, a sample was classified as aneuploid when the probability was greater than 50% and euploid when the probability was less than 50%.

Statistical analysis

The predictive performance of the machine-learning models on the primary test sets was assessed using the accuracy, 95% CI, and positive predictive value (PPV) for each model. The performance of the STORK-A deep-learning models on the primary test set was measured using accuracy, 95% CI, PPV, negative predictive value (NPV), receiver operator curves, and AUC. Sensitivity and specificity for STORK-A were reported for the models with the best performance for each classification task. The independent and external datasets, WCM-ES+ and IVI Valencia, used to assess the generalisability of STORK-A, were compared, measuring accuracy, AUC, PPV, and NPV.

To gain further insight into the specific demographic performance of STORK-A with the task of classifying aneuploids versus euploids, the primary test set was stratified across maternal ages and the day of blastocyst formation for post-hoc analysis. Predictions from the primary test set were separated into day 5 and day 6 embryos, and four age groups (<35 years; ≥35 to <37 years; ≥37 to <39 years; and ≥39 years) based on a similar procedure.[31] The accuracy for each of these demographics within the primary test were then reported. Understanding the relationship between older patients and the incidence of aneuploidy, embryos from patients aged 37–42 years in the primary test set were separated into individual groups to identify the optimal classification thresholds that maximise the sum of sensitivity and specificity. Lastly, a post-hoc analysis of fetal heart and livebirth rates was done to explore the association between STORK-A and positive outcomes in the primary test set and stratified by age group.

This study used retrospective and fully de-identified data. The study was performed in accordance with relevant guidelines and regulations, patient consent was obtained, and the study was approved by the Institutional Review Board at Weill Cornell Medicine (numbers 1401014735 and 19–06020306) and by the IVI Valencia Institutional Review Board (number 1709-VLC-094-MM).

Role of the funding source

The funding source for this study had no role in the experimental design of the study, data collection, data analysis, data interpretation, or writing of this report.

Results

Analysis and model development included the use of 10 378 embryos, all with PGT-A results, from 1385 patients. Several clinical features were used to develop predictions, including maternal age ranging from 21 to 48 years (mean 36·98 years [SD 4·62]), morphokinetic parameters, morphological assessment, and images captured at 110 h after ICSI. The race or ethnicity of patients in the dataset was primarily White (non-Hispanic) but also included Asian (non-Hispanic), Black (non-Hispanic), and Hispanic or Latinx.

Lasso regression (also known as logistic regression with L1 regularisation), was used to introduce sparsity into the model prediction and improve the interpretability of clinical features and their contributions to ploidy prediction. We found that, of these features, maternal age and blastocyst score had the greatest effect on ploidy prediction for all three tasks (appendix p 1). This result follows previously published work that shows age and blastocyst score correlated with ploidy.[28]

Additional logistic regression models were performed to provide an increasingly granular assessment of blastocyst score influence on ploidy prediction, owing to its feature importance from lasso regression. When comparing the ploidy prediction performance of a logistic regression model with ridge regression (also known as L2 regularisation), using maternal age and blastocyst score, accuracies were similar compared with when using maternal age and the three components of blastocyst grade (trophectoderm, inner cell mass, and expansion; appendix p 2). In addition, results from this model were similar to results obtained using the entire clinical feature space for complex aneuploid versus single aneuploid plus euploids, which corresponds to the high weights obtained for blastocyst score and maternal age using lasso regression. When considering feature impor tance, we find that maternal age positively correlates with aneuploid, as well as blastocyst score (appendix p 2). This result is in agreement with previous literature for maternal age, and for blastocyst score, as lower scores are defined as higher quality embryos.[32] When analysing the individual components of the blastocyst grade, we saw that changes in the trophectoderm grade had the largest effect on model performance, followed by the expansion grade, and then the inner cell mass grade. This could point to some biological relevance since the cells biopsied and whose DNA is used for sequencing are from the trophectoderm.

As an additional step to analyse feature importance, a univariate assessment of each morphological feature was done in combination with egg age. This analysis supports trophectoderm grade being most predictive of ploidy, with an accuracy of 0·703 (95% CI 0·684–0·722), followed by inner cell mass grade 0·697 (0·684–0·710), and expansion grade 0·692 (0·677–0·706) for the aneuploids versus euploids classification. This trend holds for both complex aneuploid versus euploids (0·773 [95% CI 0·766–0·780] for trophectoderm; 0·754 [0·747–0·760] for inner cell mass; and 0·743 [0·733–0·753] for expansion) and complex aneuploid versus aneuploids plus euploids (0·706 [0·694–0·717] for trophectoderm; 0·694 [0·683–0·706] for inner cell mass; and 0·682 [0·673–0·691] for expansion). A high Pearson correlation (0·84) between inner cell mass and trophectoderm grades might explain the low feature weight of the inner cell mass grade in the multivariable analysis (appendix p 3).

In terms of evaluation of machine and deep-learning models for ploidy prediction, as logistic regression is a linear model, the subsequent step in the study was to understand how increasingly complex and non-linear machine-learning approaches, specifically XGBoost, k-NN, SVM, and Random Forest, would perform when predicting embryo ploidy. Each machine-learning model was trained and tested using various combinations of clinical features across several classification tasks including aneuploids versus euploids, complex aneuploid versus single aneuploid plus euploids, and complex aneuploid versus euploids.

Across the three classification tasks, SVM and XGBoost generally performed best, except for Random Forest in the complex aneuploid versus euploids plus single aneuploid task (table 1). Among the four architectures, k-NN performed the worst. In the aneuploids versus euploids task, SVM using maternal age, morphokinetics, and blastocyst score showed an accuracy of 70·5% (95% CI 68·2–72·8%). For the complex aneuploid versus euploids plus single aneuploid task, Random Forest reported an accuracy of 76·8% (95% CI 74·6–78·9%) using maternal age, morphokinetics, and blastocyst score. Finally, in the complex aneuploid versus euploids classification task, XGBoost and SVM shared the same performance of 77·6 (95% CI 75·0–80·0%) using maternal age and blastocyst score. A review of the performance of the models with a single clinical feature indicates that maternal age alone is a strong predictor of ploidy status across all classification tasks. The addition of morphological assessments, either blastocyst grade, blastocyst score, or AIBS, to maternal age generally improved model accuracies across the board in all three classification tasks. On the other hand, morphokinetic parameters in the aneuploids versus euploids and complex aneuploid versus euploids tasks generally did not improve performance in alignment with findings from the regression analyses. However, in the complex aneuploid versus euploids plus single aneuploid task, the addition of morphokinetic parameters to Random Forest models improved performance.

Next, STORK-A (a deep-learning CNN based on a modified ResNet18 architecture) was used to extract features from static images of embryos at 110 h after ICSI that were then concatenated with the previously used clinical features to predict ploidy. At a baseline, models trained using only images for the following classification tasks reported accuracies of 59·2% (95% CI 56·7–61·6) for aneuploids versus euploids, 61·1% (58·6–63·5) for complex aneuploid versus single aneuploid plus euploids, and 64·0% (61·1–66·8) for complex aneuploid versus euploids (table 2). Similar to what was observed in the machine-learning models, the addition of morphological assessments along with maternal age improved model accuracy in all three classification tasks. Again, morphokinetic parameters did not provide substantial improvement to the models and in some cases decreased performance. The best-performing models for the aneuploids versus euploids (accuracy 69·3% [95% CI 66·9–71·5]), complex aneuploid versus euploids plus single aneuploid (74·0% [71·7–76·1]), and complex aneuploid versus euploids (77·6% [75·0–80·0]) classification tasks used images, maternal age, morphokinetic parameters, and morphological assess ment (blastocyst grade or blastocyst score). For the complex aneuploid versus euploids task, the model including image, maternal age, and blastocyst grade performed similarly with an accuracy of 77·6% (95% CI 75·1–80·1) but this resulted in a trade-off in PPV and NPV (table 2).

When tested against the primary test set, STORK-A for aneuploids versus euploids reported an accuracy of 69·3%. When the primary test aneuploids in the aneuploids versus euploids classification task were stratified it was observed that STORK-A correctly predicted 77·1% of complex aneuploid embryos and correctly predicted 57·0% of single aneuploid embryos (appendix p 3). It is plausible that single aneuploid and euploid embryos share an overlap in morphology and morphokinetics, making it difficult for STORK-A to differentiate between the two classes. STORK-A for complex aneuploid versus euploids plus single aneuploid was poised to verify this overlap assumption and reported an accuracy of 74·0%. This classifier was able to identify 89·8% of all euploid embryos, 66·7% of single aneuploid embryos, and 57·6% of complex aneuploid (appendix p 3). Given these results, it is more likely that single aneuploid embryos share an overlap among both complex aneuploid and euploids as the accuracy of the complex aneuploid class decreased, whereas euploids increased.

Our results indicate the utility of blastocyst morphology assessment, both blastocyst grade and blastocyst score, for ploidy prediction in both machine-learning and deep-learning models. However, morphology assessment is subject to observer variability and bias. To circumvent this issue, blastocyst scores were predicted utilising deep learning and regression with embryologist-derived blastocyst scores as ground truth labels. The model reported a mean squared error of 16·3 and a Pearson correlation coefficient of 0·65 for AIBS. The AIBS for each embryo in the primary dataset was then used as input for all machine-learning and deep-learning classification tasks. In general, AIBS underperformed compared with embryologist-derived blastocyst score and blastocyst grade. However, AIBS does offer an improvement over age alone in machine-learning models for all three classification tasks, and image and age alone in deep-learning models for complex aneuploid versus euploids plus single aneuploid and complex aneuploid versus euploids classification tasks.

Embryos in the primary test set and their predictions were further categorised by age groups (<35 years; ≥35 to <37 years; ≥37 to <39 years; and ≥39 years), to do a post-hoc analysis of whether there were differences in the model's ability to predict ploidy status on the basis of age. Of interest were the embryos of patients younger than 36 years and older than 39 years, as several studies have concluded that not all patients need or should use PGT-A. Within the youngest age group of the primary test set, STORK-A for aneuploids versus euploids using an image, age, morphokinetic parameters, and blastocyst score correctly classified 63·% of embryos, with a specificity of 93·5% and sensitivity of 12·0%. Although the model can sufficiently identify euploid embryos, it struggles to correctly identify aneuploids and is subject to false negatives. Therefore STORK-A might be beneficial as a screening tool to identify euploids without being hindered by a large number of false positives in embryos associated with maternal age younger than 36 years. For the embryos with maternal age older than 39 years, the same STORK-A model correctly predicted 85·1% of embryos, with a specificity of 5·4% and sensitivity of 98·5%. The severe class imbalance of aneuploids and euploids in this age group of embryos is the cause for the differences in sensitivity and specificity. Nonetheless, the high sensitivity for this age group would be useful for identifying aneuploid embryos without generating many false negative predictions.

Given these findings, an optimal threshold that maximises the sum of the specificity and sensitivity for embryos with maternal age 37–42 years was assessed in a post-hoc analysis

(appendix pp 6–7). For embryos with maternal age of 37 years, the performance using optimal threshold only slightly deviated from the baseline 50:50 threshold, and therefore did not substantially improve performance. For embryos with maternal age 38–42 years, we identified a trend that suggests optimal thresholds might be useful depending on an embryologist's intentions. That is, whether it is favourable to deselect as many aneuploids as possible but also risk the deselection of some euploids, a higher sensitivity, or confidently identifying euploid embryos but risking the inclusion of some aneuploids. For example, in embryos with a maternal age of 43 years, a 50:50 decision threshold has an accuracy of 90·9%, with a high sensitivity of 98·9%. However, this comes at the cost of predicting only one embryo as euploid, resulting in a low specificity of 11·1%. By applying an optimal decision threshold of 0·350 to maximise the sum of specificity and sensitivity, the overall accuracy drops to 70·7%, but in this instance there is a benefit, as doing so incurs an increase in the specificity that reaches 66·7% by correctly classifying 6 of 9 euploids, but misclassifying 26 aneuploids as euploid.

A downstream event to ploidy prediction is the presence of a fetal heart and livebirth. Table 3 shows the fetal heart rate and livebirth rate of 242 transferred embryos that were classified as euploid by PGT-A. In a post-hoc analysis, the ability of STORK-A for aneuploids versus euploids to correctly predict euploid embryos was compared with the fetal heart and livebirth rates of embryos determined to be euploid by PGT-A. Of the 242 embryos, STORK-A predicted 166 (69%) embryos to be euploid. Of these 166 euploid embryos, 93 (56%) resulted in a fetal heart, which was similar to the rate established by PGT-A (59%). When investigating the livebirth rates, embryos predicted to be euploid by STORK-A showed a livebirth rate of 48%, again similar to the rate observed by PGT-A (49%). For patients aged 37 years and younger, STORK-A shows an ability to correctly predict embryos that will result in fetal hearts and livebirths.

To test the robustness and generalisability of STORK-A, performance metrics from the primary test set were compared with those of two independent and external test sets. The first independent dataset, WCM-ES+, was from the Weill Cornell Medicine Center of Reproductive Medicine and included images captured using the EmbryoScope+. The dataset included 841 embryos, along with maternal age, morphokinetic parameters, and morphological assessments (blastocyst grade and blastocyst score). The second independent dataset, IVI Valencia, included images from 554 embryos captured using the original EmbryoScope. The clinical infor mation available included maternal age and morphokinetics. The trained STORK-A aneuploids versus euploids classifier (which used images, morphokinetic parameters, and maternal age) was tested on the WCM-ES+ and IVI Valencia test datasets. STORK-A produced accuracies of 63·4% (AUC 0·702) for WCM-ES+ and 65·7 (AUC 0·715) for IVI Valencia (figure 2). Compared with the accuracy of the primary test set (67·8%; AUC 0·737), we saw that STORK-A was able to maintain generalisability against the two external test sets. STORK-A for complex aneuploid versus euploids plus single aneuploid (where image, maternal age, morphokinetics parameters, and blastocyst grade were used as inputs) was tested on the WCM-ES+ test data and resulted in an accuracy of 74·7% (AUC 0·781), similar to the accuracy for the primary test set, which was 74·0% (AUC 0·760; appendix p 5).

As part of this work, we developed a user-friendly web-based app for STORK-A (figure 3). The platform requires, at a minimum, an image of a blastocyst. Users then have the option to include patient age, morphological assessment, and complete morphokinetic parameters from time from pro-nuclear fading to time of the start of blastulation. The results include probabilities for each of the three classifiers.

Discussion

In this study, it was observed that among the most important feature for ploidy classification was maternal age at the time of oocyte retrieval, which is known to correlate with the incidence of aneuploidy. Additionally, morphological features had an important role in ploidy prediction as an embryologist-derived improved model performance. Conversely, morphokinetics were found to have a less important role in classification. The median impu tation of missing morphokinetic parameters did not alter these results, which we verified by retraining all machine and deep-learning models (appendix pp 8–10).

Several studies have attempted to mimic the skill and experience of trained embryologists in assessing embryo quality while simultaneously improving consistency and reducing bias through the development of unbiased and automated embryo selection tools using deep learning. STORK, an embryo morphological assessment model based on the Veeck and Zaninovic grading system,[27] used transfer learning with Inception-v1.[9] STORK predicted embryo quality with near-perfect accuracy and showed that its good-quality predictions were associated with better livebirth outcomes. A similar study aimed to automate embryo grading using deep learning on a ResNet50 architecture.[33] A model for rank-based selection of embryos was developed on the basis of quality in addition to assessing the implantation potential of embryos.[34] Although automated morphological assessment is useful for developing a standardised method of grading embryo quality, these methods do not address the need to non-invasively predict the ploidy status of embryos as a means of prioritising and selecting embryos with the highest implantation potential.

Our deep-learning approach, STORK-A, showed an ability to classify the ploidy status of embryos in three distinct classification tasks: aneuploids versus euploids; complex aneuploid versus euploids plus single aneuploid; and complex aneuploid versus euploids. The best models of these three classifiers incorporated an image, maternal age (age at the time of oocyte retrieval), morphokinetic parameters, and morphological assessment (blastocyst grade or blastocyst score). Overall, the use of static images of embryos at 110 h after ICSI to predict ploidy status did not markedly improve the performance of STORK-A when comparing machine-learning models and deep-learning models across three classification tasks. This might be due to the images capturing embryos in different stages of develop ment at 110 h after ICSI, in which case the deep-learning models are learning to distinguish features that differ between morulas and blastocysts, rather than differences only between blastocysts, as is the case for morphological assess ment. We also show that a standardised deep-learning regression model to predict AIBS offers an improvement over age alone in the machine-learning and deep-learning models. However, we note that the performance gain is not as high when compared with using embryologist-derived morpho logical assessment. Nonetheless, we show as a proof of concept that artificial intelligence can approach the

performance of expert annotators in providing useful information about embryo implantation potential. We hypothesise that the inclusion of temporal and spatial information from videos of embryo development, rather than a single static image, might yield improved ploidy predictive performance.

Several limitations arose in the study. First, embryos in the dataset used to train, validate, and test STORK-A were previously selected by embryologists as candidates for PGT-A on the basis of their morphology. Those embryos that were not biopsied for PGT-A were therefore not included. In other studies, unused embryos have been included and labelled as negative results. This work deviated from other studies to gain increased confidence in STORK-A's ability to detect ploidy; however, this does have the potential to bias the dataset. An ideal dataset would include PGT-A results for all embryos regardless of their morphological quality.

Another limitation was the use of images captured only by time-lapse microscopy, thereby limiting generalisability. Time-lapse machines are costly, and few clinics use this technology. However, because STORK-A makes use of single static images, future development will incorporate images of embryos captured using different imaging modalities from different clinics, which will improve generalisability.

Morphokinetic annotations (which require time-lapse machinery) and morphological assessments were incorporated into STORK-A, thereby introducing human bias, which could limit generalisability. An ideal artificial-intelligence model would not be trained using unstandardised and subjective observations such as morpho kinetic parameters and morphological assess ment; instead, it would be trained on standardised and reproducible data. To address this limitation, we attempted to use an artificial-intelligence-driven predicted blastocyst score (AIBS), which is standardised and reproducible. However, it should be noted that the accuracies of STORK-A classifiers that use an image and maternal age only show decreases of 2–4% compared with classifiers that incorporate subjective morphokinetic parameters and morphological assessments, and were generalisable to the independent test sets (appendix p 5).

Finally, next-generation sequencing can distinguish euploids, several types of aneuploids, and high-level or low-level mosaic embryos. However, differences in mosaic reporting across genetic laboratories introduce limitations to generalisability when these results are used as ground truth labels. In this study, PGT-A results for the primary dataset from Weill Cornell Medicine and the WCM-ES+ independent dataset were generated from the same genetics laboratory. Between both datasets, 719 embryos had detailed genetic information and, of those, only 32 (4%) were mosaic and categorised as euploid. The IVI Valencia independent dataset included PGT-A results from a different genetics laboratory that did not provide detailed sequencing information of embryos, and instead determined embryos to be euploid or aneuploid. Because of the narrow reporting of sequencing information, mosaicism was not considered during model development and therefore cannot be assessed by STORK-A. Therefore, the binary classification scheme of STORK-A introduces a limitation, as mosaic embryos with high-implantation potential could be misclassified.

This study suggests a future role for STORK-A in the fertility clinic. However, STORK-A in its current state is not intended to be a derivative of a prenatal test or replace PGT-A, as further development and randomised clinical studies would be required beforehand. Rather, STORK-A is intended to be an assistive decision-making tool that provides a standardised, non-invasive, and cost-efficient means of selecting and prioritising high-quality embryos for PGT-A biopsy or transfer to patients, as opposed to using traditional methods such as morphological assess ment, which are biased and subjective. STORK-A for complex aneuploid versus euploids plus single aneuploid in particular would be very beneficial in the clinic. The high specificity of 80·1% could assist in identifying euploid and single aneuploid embryos without misclassifying a large number of complex aneuploid when prioritising embryos for biopsy or transfer. An embryologist could confidently assess an embryo as truly being euploid or single aneuploid with a negative predictive value of 82·3%.

The question about the actual benefit of PGT-A is pertinent to this study. Although PGT-A can detect chromosomal abnormalities with great accuracy, a Cochrane review[35] found that there is insufficient evidence to support its clinical use, as it has not led to increased pregnancy rates or livebirth outcomes. If the ultimate goal of developing assistive reproductive technologies for IVF is to reduce a patient's time to pregnancy and improved livebirth outcomes, that should be our gold standard. However, as it stands, embryo selection is still crucial to this outcome, and embryos with the greatest implantation potential must be selected or prioritised, whereas those with low-implantation potential are deprioritised. Current widespread screening methods such as morphological assessments for embryo selection are unstandardised, and subjective, except for PGT-A. Standardisation, free of variability, is necessary for the development of methods to prioritise and select embryos that are consistent across clinics. For this reason, we elected to use PGT-A results as the ground truth labels for the development of our models. STORK-A is poised to provide standardised embryo selection and prioritisation in a manner that is noninvasive, cost-efficient, and time-efficient.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

IH is supported by a US National Institute of General Medical Sciences Maximizing Investigators' Research Award (grant number R35 GM138152-01). Through allocation number TG-ASC190055 to IH, this study used the Extreme Science and Engineering Discovery Environment, supported by the US National Science Foundation (grant number ACI-1548562). JB is supported by a Weill Cornell Medicine Clinical & Translations Science Center training award (US National Institutes of Health–US National Centre for Advancing Translational Sciences grant number TL1-TR-002386).

OE is scientific adviser for, and an equity holder in, Freenome, Owkin, Volastra Therapeutics, OneThree Biotech, Genetic Intelligence, Acuamark DX, Harmonic Discovery, and Champions Oncology, and has received funding from Eli Lilly, Johnson & Johnson–Janssen, Sanofi, AstraZeneca, and Volastra. NZ is a paid consultant for AIVF and Fairtility, and is on the advisory board of, and has equity in, Alife Health. IH gave an academic lecture for Fairtility on a related topic (precision medicine and artificial intelligence: what we have learned and how it can impact assisted reproductive technology). JB, JEM, ZR, OE, NZ, and IH are listed as inventors on a provisional patent filed by Cornell University (application number 63/308,710) about the technology described in this study. MM received speaker fees from Merck, Vitrolife, Ferring, Theramex, and Gideon Richter. PZ holds stocks in Pfizer

and Bristol Myers Squibb. JB received funding support for attending meetings for the 2022 Association for Clinical and Translational Science Conference.

References

1. Herbert M, Kalleas D, Cooney D, Lamb M, Lister L. Meiosis and maternal aging: insights from aneuploid oocytes and trisomy births. Cold Spring Harb Perspect Biol 2015; 7: a017970.
2. Gardner DK, Sakkas D. Assessment of embryo viability: the ability to select a single embryo for transfer—a review. Placenta 2003; 24 (suppl B): S5–12. [PubMed: 14559024]
3. Gardner DK, Meseguer M, Rubio C, Treff NR. Diagnosis of human preimplantation embryo viability. Hum Reprod Update 2015; 21: 727–47. [PubMed: 25567750]
4. Meseguer M, Rubio I, Cruz M, Basile N, Marcos J, Requena A. Embryo incubation and selection in a time-lapse monitoring system improves pregnancy outcome compared with a standard incubator: a retrospective cohort study. Fertil Steril 2012; 98: 1481–9.e10. [PubMed: 22975113]
5. Storr A, Venetis CA, Cooke S, Kilani S, Ledger W. Inter-observer and intra-observer agreement between embryologists during selection of a single day 5 embryo for transfer: a multicenter study. Hum Reprod 2017; 32: 307–14. [PubMed: 28031323]
6. Tunis SR, Clarke M, Gorst SL, et al. Improving the relevance and consistency of outcomes in comparative effectiveness research. J Comp Eff Res 2016; 5: 193–205. [PubMed: 26930385]
7. Paternot G, Devroe J, Debrock S, D'Hooghe TM, Spiessens C. Intra- and inter-observer analysis in the morphological assessment of early-stage embryos. Reprod Biol Endocrinol 2009; 7: 105. [PubMed: 19788739]
8. Sundvall L, Ingerslev HJ, Breth Knudsen U, Kirkegaard K. Inter- and intra-observer variability of time-lapse annotations. Hum Reprod 2013; 28: 3215–21. [PubMed: 24070998]
9. Khosravi P, Kazemi E, Zhan Q, et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. NPJ Digit Med 2019; 2: 21. [PubMed: 31304368]
10. Lee H-L, McCulloh DH, Hodes-Wertz B, Adler A, McCaffrey C, Grifo JA. In vitro fertilization with preimplantation genetic screening improves implantation and live birth in women age 40 through 43. J Assist Reprod Genet 2015; 32: 435–44. [PubMed: 25578536]
11. Simon AL, Kiehl M, Fischer E, et al. Pregnancy outcomes from more than 1,800 in vitro fertilization cycles with the use of 24-chromosome single-nucleotide polymorphism-based preimplantation genetic testing for aneuploidy. Fertil Steril 2018; 110: 113–21. [PubMed: 29908770]
12. Munné S, Kaplan B, Frattarelli JL, et al. Preimplantation genetic testing for aneuploidy versus morphology as selection criteria for single frozen-thawed embryo transfer in good-prognosis patients: a multicenter randomized clinical trial. Fertil Steril 2019; 112: 1071–1079.e7. [PubMed: 31551155]
13. Harper JC, Geraedts J, Borry P, et al. Current issues in medically assisted reproduction and genetics in Europe: research, clinical practice, ethics, legal issues and policy. Eur J Hum Genet 2013; 21 (suppl 2): S1–21.
14. Xu J, Fang R, Chen L, et al. Noninvasive chromosome screening of human embryos by genome sequencing of embryo culture medium for in vitro fertilization. Proc Natl Acad Sci USA 2016; 113: 11907–12. [PubMed: 27688762]
15. Huang L, Bogale B, Tang Y, Lu S, Xie XS, Racowsky C. Noninvasive preimplantation genetic testing for aneuploidy in spent medium may be more reliable than trophectoderm biopsy. Proc Natl Acad Sci USA 2019; 116: 14105–12. [PubMed: 31235575]
16. Cimadomo D, Capalbo A, Ubaldi FM, et al. The impact of biopsy on human embryo developmental potential during preimplantation genetic diagnosis. Biomed Res Int 2016; 2016: 7193075. [PubMed: 26942198]
17. Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Hickman CFL. Modelling a risk classification of aneuploidy in human embryos using non-invasive morphokinetics. Reprod Biomed Online 2013; 26: 477–85. [PubMed: 23518033]

18. Basile N, Nogales MC, Bronet F, et al. Increasing the probability of selecting chromosomally normal embryos by time-lapse morphokinetics analysis. Fertil Steril 2014; 101: 699–704. [PubMed: 24424365]
19. Kramer YG, Kofinas JD, Melzer K, et al. Assessing morphokinetic parameters via time lapse microscopy (fTLM) to predict euploidy: are aneuploidy risk classification models universal? J Assist Reprod Genet 2014; 31: 1231–42. [PubMed: 24962789]
20. Chawla M, Fakih M, Shunnar A, et al. Morphokinetic analysis of cleavage stage embryos and its relationship to aneuploidy in a retrospective time-lapse imaging study. J Assist Reprod Genet 2015; 32: 69–75. [PubMed: 25395178]
21. Minasi MG, Colasante A, Riccio T, et al. Correlation between aneuploidy, standard morphology evaluation and morphokinetic development in 1730 biopsied blastocysts: a consecutive case series study. Hum Reprod 2016; 31: 2245–54. [PubMed: 27591227]
22. Patel DV, Shah PB, Kotdawala AP, Herrero J, Rubio I, Banker MR. Morphokinetic behavior of euploid and aneuploid embryos analyzed by time-lapse in embryoscope. J Hum Reprod Sci 2016; 9: 112–18. [PubMed: 27382237]
23. Mumusoglu S, Yarali I, Bozdag G, et al. Time-lapse morphokinetic assessment has low to moderate ability to predict euploidy when patient- and ovarian stimulation-related factors are taken into account with the use of clustered data analysis. Fertil Steril 2017; 107: 413–21.e4. [PubMed: 27939508]
24. Del Carmen Nogales M, Bronet F, Basile N, et al. Type of chromosome abnormality affects embryo morphology dynamics. Fertil Steril 2017; 107: 229–35.e2. [PubMed: 27816230]
25. Chavez-Badiola A, Flores-Saiffe-Farías A, Mendizabal-Ruiz G, Drakeley AJ, Cohen J. Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. Reprod Biomed Online 2020; 41: 585–93. [PubMed: 32843306]
26. Lee C-I, Su Y-R, Chen C-H, et al. End-to-end deep learning for recognition of ploidy status using time-lapse videos. J Assist Reprod Genet 2021; 38: 1655–63. [PubMed: 34021832]
27. Veeck LL, Zaninovic N. An atlas of human blastocysts. Boca Raton, FL: CRC Press, 2019.
28. Zhan Q, Sierra ET, Malmsten J, Ye Z, Rosenwaks Z, Zaninovic N. Blastocyst score, a blastocyst quality ranking tool, is a predictor of blastocyst ploidy and implantation potential. F S Rep 2020; 1: 133–41. [PubMed: 34223229]
29. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. arXiv 2017; published online Nov 25. 10.48550/arXiv.1705.07874 (preprint).
30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition; June 27–30, 2016 (pp 770–78).
31. Irani M, Zaninovic N, Rosenwaks Z, Xu K. Does maternal age at retrieval influence the implantation potential of euploid blastocysts? Am J Obstet Gynecol 2019; 220: 379.e1–7.
32. Demko ZP, Simon AL, McCoy RC, Petrov DA, Rabinowitz M. Effects of maternal age on euploidy rates in a large cohort of embryos analyzed with 24-chromosome single-nucleotide polymorphism-based preimplantation genetic screening. Fertil Steril 2016; 105: 1307–13. [PubMed: 26868992]
33. Chen T-J, Zheng W-L, Liu C-H, Huang I, Lai H-H, Liu M. Using deep learning with large dataset of microscope images to develop an automated embryo grading system. Fertil Rep 2019; 1: 51–56.
34. Bormann CL, Kanakasabapathy MK, Thirumalaraju P, et al. Performance of a deep learning based neural network in the selection of human blastocysts for implantation. eLife 2020; 9: e55301. [PubMed: 32930094]
35. Cornelisse S, Zagers M, Kostova E, Fleischer K, van Wely M, Mastenbroek S. Preimplantation genetic testing for aneuploidies (abnormal number of chromosomes) in in vitro fertilisation. Cochrane Database Syst Rev 2020; 9: CD005291.

Research in context

Evidence before this study

The current methods of embryo selection for transfer during in-vitro fertilisation are subject to inter-observer and intra-observer bias as observed in morphological assessment and morphokinetic annotation, or present an ethical barrier as seen in invasive trophectoderm biopsies for preimplantation genetic testing for aneuploid (PGT-A). We searched PubMed and Google Scholar for articles published from Jan 1, 2000 to June 5, 2021, using the search terms [“ivf” OR “in vitro fertilization”] AND “embryo selection” AND “quality” AND “ploidy” AND “aneuploid” AND “euploid” AND [“artificial intelligence” OR “machine learning” OR “deeplearning”]. We found that several studies had sought to alleviate the shortcomings of morphological assessment by using deep-learning approaches to predict embryo quality. However, few studies have attempted to use deep learning to predict embryo ploidy status as a standardised method of embryo selection.

Added value of this study

STORK-A was designed to non-invasively predict embryo ploidy. Using a dataset with images at 110 h after intracytoplasmic sperm injection and clinical information for 10 378 embryos, several machine-learning and deep-learning models were developed to evaluate which features contribute to ploidy classification. Maternal age, along with morphological assessment, were strong predictors of embryo ploidy, while morphokinetic parameters did not contribute to improving predictions.

Implications of all the available evidence

STORK-A is not intended to replace PGT-A. Instead, it demonstrates a strong ability to correctly predict euploid and single aneuploid embryos which could be used to supplement traditional methods of embryo selection and prioritisation. This study also shows the generalisability of STORK-A via the testing of independent datasets. Lastly, single static images at 110 h after intracytoplasmic sperm injection alone are not sufficient for predicting embryo ploidy. Models built using videos of embryo development with both spatial and temporal information are likely to provide improved predictive ability.

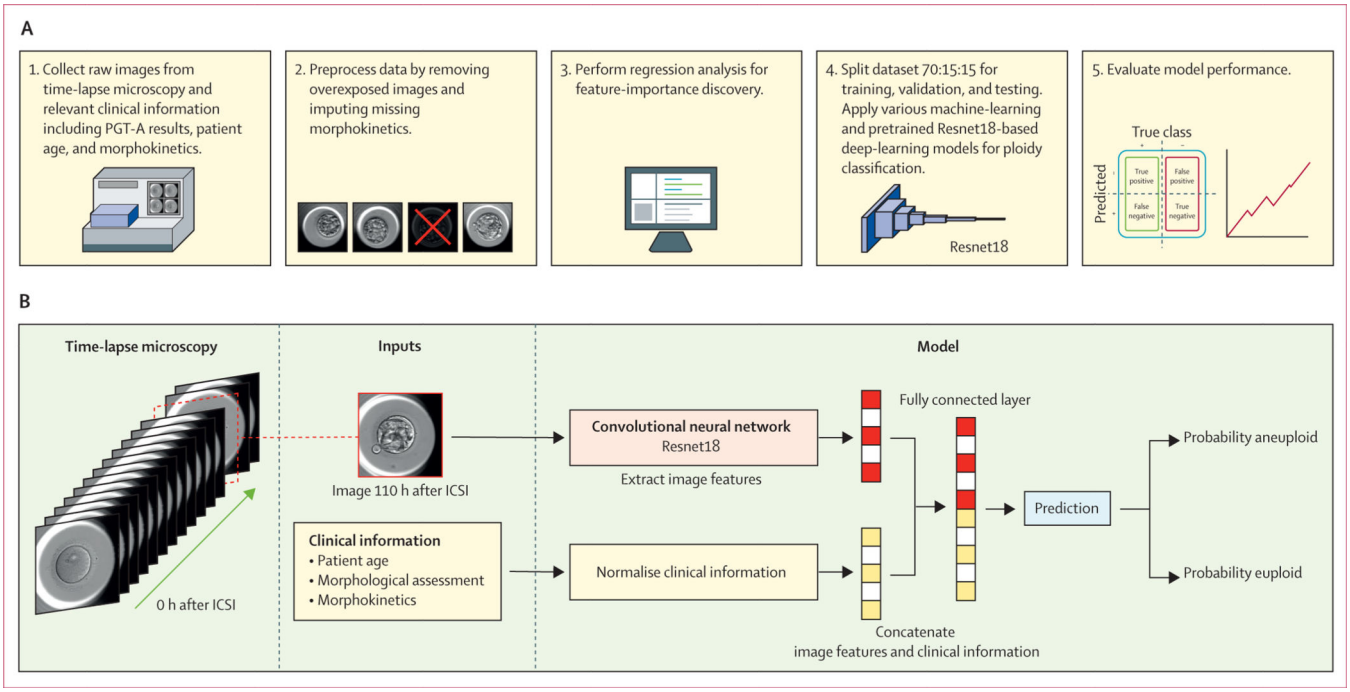

Figure 1: Study design and STORK-A schematic

(A) First, time-lapse videos are extracted from the Embryoscope, and a single static image at 110 h after ICSI (focal plane 0) is used for each embryo, along with morphokinetic annotations, morphological assessments, maternal age, and associated PGT-A results. Next, the dataset is preprocessed to remove underexposed images by manual detection, and missing morphokinetic values are imputed using median imputation. Lasso and logistic regressions are then applied to clinical information to determine feature importance. After determining feature importance, the dataset is split 70:15:15 for training, validating, and testing models to predict embryo ploidy. Hyperparameters for the models are then optimised through iterative training and once completed, the performance on the test set is evaluated. (B) Overview of our proposed deep-learning model for ploidy classification. Image features extracted from the ResNet18 convolutional neural network are concatenated with clinical information (maternal age, morphokinetic parameters, and one of three morphological assessments: blastocyst grade, blastocyst score, artificial intelligence blastocyst scores) before being passed on to a final fully connected layer. ICSI=intracytoplasmic sperm injection. PGT-A=preimplantation genetic testing for aneuploidy.

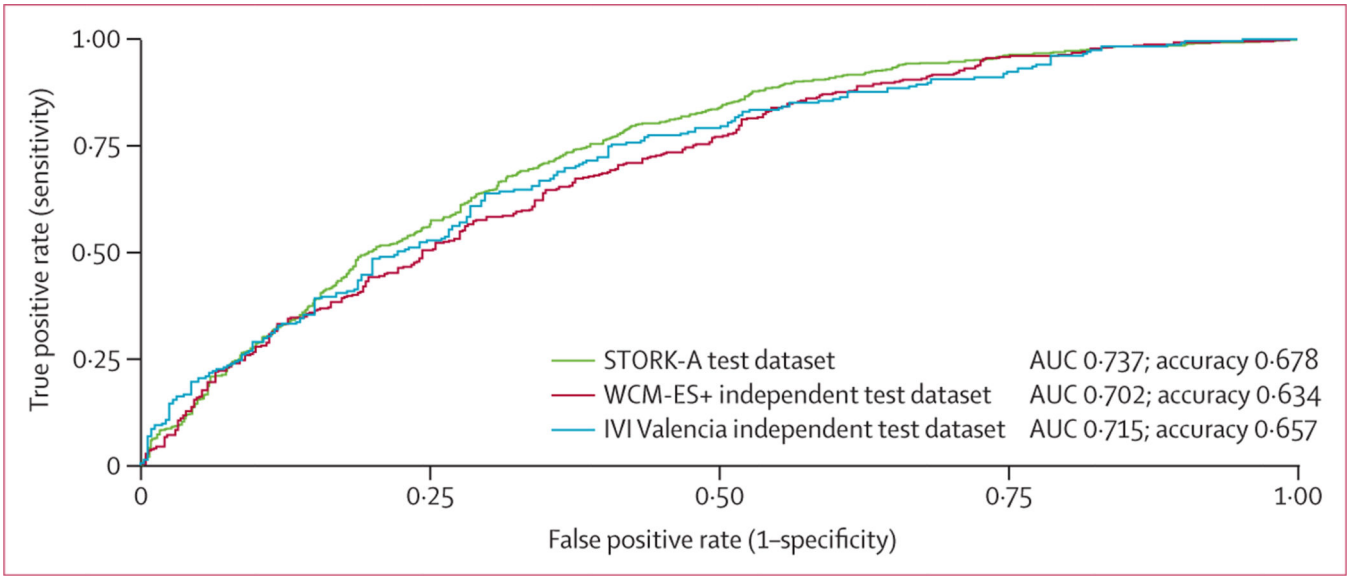

Figure 2: Independent dataset validation

Trained STORK-A for aneuploids versus euploids classification using images, maternal age, and morphokinetic parameters reported similar accuracies on the IVI Valencia test dataset and WCM-ES+ test dataset when compared with the STORK-A primary test dataset. AUC= area under the receiver operating characteristic curve.

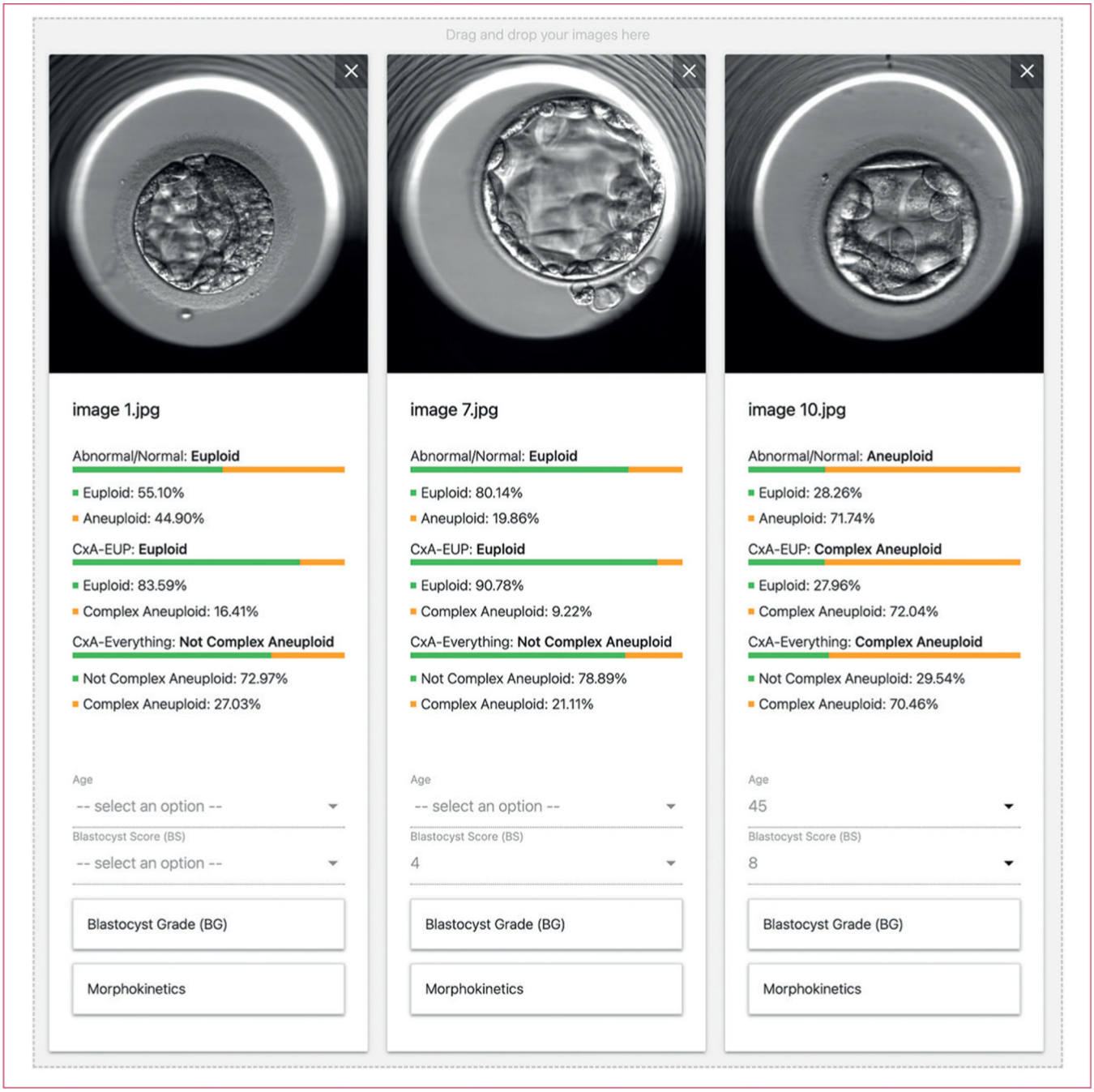

Figure 3: STORK-A web interface

An automated platform that can be used in clinical settings to evaluate ploidy status as a support tool for embryologists.

Table 1:

Machine-learning performance

	XGBoost		k-NN		SVM		Random Forest	
	Accuracy (95% CI)	**PPV**	**Accuracy (95% CI**	**PPV**	**Accuracy (95% CI**	**PPV**	**Accuracy (95% CI**	**PPV**
Aneuploids versus euploids								
Age	66·3% (63·9–68·7)	74·1%	66·3% (63·9–68·7)	74·1%	66·3% (63·9–68·7)	74·1%	66·3% (63·9–68·7)	68·7%
Morphokinetics	54·8% (52·3–57·3)	62·3%	52·5% (50·0–55·0)	59·9%	55·4% (52·9–57·9)	64·0%	56·5% (54·0–59·0)	61·8%
Blastocyst score	62·7% (60·2–65·1)	74·0%	NA	NA	62·7% (60·2–65·1)	74·0%	62·7% (60·2–65·1)	74·0%
Blastocyst grade	63·2% (60·8–65·6)	72·2%	61·4% (58·9–63·8)	72·6%	61·0% (58·5–63·4)	71·1%	62·4% (59·9–64·8)	74·5%
AIBS	57·5% (55·0–60·0)	64·7%	56·5% (54·0–59·0)	63·4%	57·4% (54·9–59·8)	64·9%	53·7% (51·2–56·2)	59·4%
Age plus morphokinetics	65·3% (62·8–67·6)	72·6%	63·6% (61·2–66·0)	71·1%	66·0% (63·6–68·4)	73·5%	65·1% (62·7–67·5)	70·0%
Age plus blastocyst score	68·5% (66·1–70·8)	78·0%	69·7% (67·4–72·0)	77·5%	70·2% (67·8–72·4)	77·8%	69·1% (66·8–71·4)	77·8%
Age plus blastocyst grade	68·5% (66·1–70·8)	77·3%	66·4% (64·0–68·8)	75·6%	69·0% (66·6–71·3)	76·9%	67·3% (64·9–69·7)	75·8%
Age plus AIBS	66·8% (64·4–69·2)	74·6%	63·4% (61·0–65·8)	70·6%	67·7% (65·3–70·1)	70·0%	61·2% (58·7–63·7)	66·4%
Age plus morphokinetics plus blastocyst score	68·1% (65·7–70·4)	75·4%	63·1% (60·6–65·5)	72·2%	70·5% (68·2–72·8)	78·6%	67·5% (65·1–69·8)	73·2%
Age plus morphokinetics plus blastocyst grade	67·4% (65·0%–69·7)	73·1%	64·3% (61·9–66·7)	72·1%	69·2% (66·8–71·5)	77·9%	68·6% (66·2–70·9)	70·9%
Age plus morphokinetics plus AIBS	65·7% (63·3–68·1)	71·9%	61·7% (59·2–64·1)	68·5%	66·2% (63·8–68·5)	73·9%	66·8% (64·4–69·2)	71·4%
Complex aneuploids versus euploids plus single aneuploids								
Age	68·1% (65·7–70·4)	46·7%	NA	NA	71·1% (68·8–73·4)	50·2%	68·1% (65·7–70·4)	46·7%
Morphokinetics	60·6% (58·1–63·0)	33·3%	53·6% (51·1–56·1)	31·8%	59·5% (57·0–61·9)	37·4%	68·9% (66·6–71·2)	42·1%
Blastocyst score	66·0% (63·6–68·4)	44·1%	NA	NA		44·1%	66·0% (63·6–68·4)	44·1%
Blastocyst grade	65·3% (62·8–67·6)	43·2%	NA	NA	64·6% (62·2–67·0)	42·8%	65·2% (62·8–67·6)	43·4%
AIBS	57·4% (54·9–59·8)	36·1%	55·9% (53·4–58·4)	34·1%	59·0% (56·5–61·5)	38·5%	60·8% (58·4–63·3)	31·9%
Age plus morphokinetics	69·9% (67·6–72·2)	48·2%	64·1% (61·7–66·5)	42·1%	70·0% (67·7–72·3)	48·8%	75·2% (73·0–77·3)	58·9%
Age plus blastocyst score	71·6% (69·3–73·8)	50·7%	71·6% (69·3–73·8)	50·8%	72·5% (70·2–74·7)	52·0%	71·4% (69·1–73·6)	50·5%
Age plus blastocyst grade	71·3% (69·0–73·6)	50·4%	69·6% (67·3–71·9)	48·4%	72·2% (69·9–74·4)	51·6%	70·5% (68·2–72·8)	49·4%
Age plus AIBS	70·2% (67·9–72·5)	49·0%	65·7% (63·3–68·1)	43·9%	71·0% (68·7–73·2)	50·0%	68·3% (65·9–70·6)	45·1%
Age plus morphokinetics plus blastocyst score	72·2% (69·9–74·4)	51·8%	66·2% (63·8–68·6)	44·3%	72·8% (70·5–75·0)	52·4%	76·8% (74·6–78·9)	61·8%
Age plus morphokinetics plus blastocyst grade	72·0% (69·7–74·2)	51·4%	65·0% (62·5–67·3)	42·9%	72·0% (69·7–74·2)	51·3%	75·2% (73·0–77·3)	57·9%
Age plus morphokinetics plus AIBS	71·0%% (68·7–73·2)	50·0%	65·0%% (62·6–67·4)	43·3%	70·8% (68·5–73·1)	49·8%	74·8% (72·6–76·9)	58·0%
Complex aneuploids versus euploids								

	XGBoost		k-NN		SVM		Random Forest	
	Accuracy (95% CI)	**PPV**	**Accuracy (95% CI**	**PPV**	**Accuracy (95% CI**	**PPV**	**Accuracy (95% CI**	**PPV**
Age	74·1% (71·5–76·7)	67·4%	74·1% (71·5–76·7)	67·4%	74·1% (71·4–76·6)	67·2%	74·1% (71·4–76·6)	67·2%
Morphokinetics	59·7% (56·7–62·6)	50·2%	53·6% (50·6–56·6)	44·1%	61·3% (58·4–64·2)	51·9%	59·2% (56·2–62·1)	49·3%
Blastocyst score	72·3% (69·5–74·9)	65·8%	72·3% (69·5–74·9)	65·8%	72·3% (69·5–74·9)	65·8%	72·3% (69·5–74·9)	65·8%
Blastocyst grade	70·6% (67·9–73·3)	63·2%	70·8% (68·1–73·5)	63·4%	71·2% (68·4–73·8)	63·5%	71·6% (68·9–74·3)	65·4%
AIBS	63·3% (60·4–66·1)	53·6%	60·4% (57·5–63·3)	63·3%	63·6% (60·7–66·5)	54·1%	57·4% (54·4–60·3)	47·3%
Age plus morphokinetics	72·6% (69·9–75·2)	67·0%	68·8% (65·9–71·5)	60·2%	73·7% (71·0–76·3)	67·0%	72·6% (69·9–75·2)	68·6%
Age plus blastocyst score	77·6% (75·0–80·0)	73·0%	77·4% (74·8–79·8)	73·2%	77·6% (75·0–80·0)	73·1%	77·0%% (74·4–79·3)	72·9%
Age plus blastocyst grade	76·8% (74·2–79·3)	71·9%	76·7% (74·1–79·1)	71·3%	76·9% (74·3–79·4)	71·3%	76·6% (74·0–79·0)	71·8%
Age plus AIBS	75·5% (72·9–78·0)	69·2%	72·6% (69·9–75·2)	65·3%	75·2% (72·6–77·7)	69·1%	68·9% (66·0–71·6)	61·6%
Age plus morphokinetics plus blastocyst score	76·5% (73·9–78·9)	71·5%	75·0%% (72·3–77·5)	67·7%	77·3% (74·7–79·7)	72·1%	76·2% (73·6–78·7)	73·0%
Age plus morphokinetics plus blastocyst grade	77·2% (74·6–79·6)	72·5%	74·1% (71·5–76·7)	67·0%	77·2% (74·6–79·6)	71·9%	75·9% (73·2–78·3)	71·5%
Age plus morphokinetics plus AIBS	73·1% (70·4–75·7)	67·5%	70·4% (67·6–73·0)	61·8%	75·6% (73·0–78·1)	69·6%	74·1% (71·4–76·6)	70·4%

Data are % (95% CI) or %. Performance measurements include accuracy (95% CI), and PPV (%). Four machine-learning architectures, XGBoost, k-NN, SVM, and Random Forest algorithms were trained, validated, and tested for three classification tasks, aneuploids versus euploids, complex aneuploids versus single aneuploids plus euploids, and complex aneuploids versus euploids. Various combinations of clinical features including maternal age, morphological assessment including blastocyst score, blastocyst grade, and AIBS, and morphokinetic parameters, were used for input. Performance measurements include percentage accuracy (95% CI), and PPV. AIBS=artificial intelligence blastocyst scores. k-NN=k-nearest neighbours. NA=not available because of an error in training owing to numerous ties in k-NN. PPV=positive predictive value. SVM=support vector machine. XGBoost=extreme gradient boost decision tree.

Table 2:

Deep-learning performance

	Deep-learning performance			Stratified test data accuracy					
	Model accuracy (95% CI)	PPV	NPV	Maternal age <35 years	Maternal age ≥35 to <37 years	Maternal age ≥37 to <39 years	Maternal age ≥39 years	Day 5	Day 6
Aneuploids versus euploids									
Image	59·2% (56·7–61·6)	63·7%	52·3%	52·1%	56·3%	62·5%	67·1%	55·7%	61·3%
Image plus morphokinetics	58·3% (55·8–60·8)	60·3%	52·1%	45·2%	52·4%	62·2%	75·6%	49·8%	63·6%
Image plus blastocyst grade	63·9% (61·5–66·3)	71·2%	56·6%	61·2%	61·4%	63·0%	70·2%	57·6%	67·8%
Image plus blastocyst score	62·2% (59·7–64·6)	69·2%	54·9%	60·4%	57·6%	60·6%	69·7%	54·7%	66·8%
Image plus AIBS	59·3% (56·8–61·7)	62·9%	52·7%	52·3%	55·3%	60·9%	69·7%	53·7%	62·7%
Image plus age	67·8% (65·5–70·2)	73·4%	61·3%	63·7%	55·3%	64·9%	85·9%	66·6%	68·6%
Image plus age plus morphokinetics	67·8% (65·4–70·1)	73·5%	61·2%	63·5%	56·3%	64·3%	85·6%	66·0%	68·8%
Image plus age plus blastocyst grade	68·7% (66·3–71·0)	71·5%	64·4%	63·5%	57·2%	67·3%	85·6%	65·2%	70·8%
Image plus age plus blastocyst score	69·0% (66·6–71·3)	75·1%	62·3%	65·1%	59·2%	64·9%	85·6%	65·4%	71·2%
Image plus age plus AIBS	66·5% (64·1–68·8)	71·2%	60·4%	61·8%	54·7%	63·8%	84·3%	65·0%	67·4%
Image plus age plus blastocyst grade plus morphokinetics	68·9% (66·5–71·2)	74·2%	62·6%	63·9%	60·1%	66·0%	84·8%	66·7%	70·2%
Image plus age plus blastocyst score plus morphokinetics	69·3% (66·9–71·5)	76·1%	62·1%	63·9%	62·1%	65·7%	85·1%	66·2%	71·1%
Image plus age plus AIBS plus morphokinetics	67·0% (64·6–69·3)	71·6%	61·1%	64·5%	51·4%	64·1%	85·3%	65·2%	68·1%
Complex aneuploids versus euploids plus single aneuploids									
Image	61·1% (58·6–63·5)	37·0%	76·0%	67·1%	60·2%	61·0%	55·0%	75·9%	52·2%
Image plus morphokinetics	61·5% (59·0–63·9)	37·1%	75·7%	68·4%	63·1%	59·4%	54·0%	76·9%	52·2%
Image plus blastocyst grade	67·1% (64·7–69·4)	44·5%	79·7%	73·1%	67·8%	63·4%	63·0%	73·3%	63·3%
Image plus blastocyst score	68·2% (65·8–70·5)	45·5%	78·6%	75·6%	69·9%	64·2%	62·0%	76·2%	63·3%
Image plus AIBS	64·4% (62·0–66·8)	38·8%	75·1%	71·8%	69·6%	62·6%	53·2%	76·7%	57·1%
Image plus age	70·5% (68·1–72·7)	48·8%	76·5%	85·8%	75·2%	61·5%	57·3%	77·2%	66·4%
Image plus age plus morphokinetics	71·6% (69·3–73·8)	50·8%	84·1%	87·6%	79·9%	57·8%	59·1%	76·4%	68·7%
Image plus age plus blastocyst grade	72·2% (69·9–74·4)	51·7%	83·5%	87·3%	77·0%	60·5%	62·0%	77·1%	69·3%
Image plus age plus blastocyst score	73·2% (70·9–75·4)	53·2%	83·4%	86·2%	75·8%	62·6%	66·1%	79·3%	69·5%
Image plus age plus AIBS	72·0% (69·7–74·2)	51·4%	83·0%	86·2%	79·1%	59·2%	61·7%	78·8%	67·9%

	Deep-learning performance			Stratified test data accuracy					
	Model accuracy (95% CI)	PPV	NPV	Maternal age <35 years	Maternal age ≥35 to <37 years	Maternal age ≥37 to <39 years	Maternal age ≥39 years	Day 5	Day 6
Image plus age plus blastocyst grade plus morphokinetics	74·0% (71·7–76·1)	54·9%	82·3%	87·6%	77·3%	62·6%	66·3%	78·8%	71·1%
Image plus age plus blastocyst score plus morphokinetics	71·6% (69·3–73·8)	50·8%	83·9%	87·1%	72·9%	61·5%	62·2%	77·4%	68·1%
Image plus age plus AIBS plus morphokinetics	73·0% (70·7–75·2)	53·3%	81·4%	86·9%	80·2%	62·1%	61·2%	77·6%	70·2%
Complex aneuploids versus euploids									
Image	64·0% (61·1–66·8)	55·2%	70·7%	67·6%	69·3%	58·5%	60·9%	74·4%	58·1%
Image plus morphokinetics	62·6% (59·7–65·4)	54·4%	66·9%	70·6%	68·0%	55·1%	56·1%	76·6%	54·4%
Image plus blastocyst grade	71·5% (68·7–74·1)	67·0%	73·9%	75·5%	73·8%	69·1%	67·3%	77·1%	68·2%
Image plus blastocyst score	70·6% (67·9–73·3)	64·5%	74·5%	74·5%	71·1%	69·4%	67·0%	76·1%	67·5%
Image plus AIBS	63·3% (60·4–66·1)	55·1%	68·1%	70·6%	70·2%	60·0%	52·7%	76·1%	55·9%
Image plus age	75·0% (72·4–77·6)	68·1%	80·2%	87·0%	72·9%	58·9%	77·9%	81·0%	71·6%
Image plus age plus morphokinetics	75·3% (72·7–77·8)	69·3%	79·5%	87·3%	72·0%	61·5%	76·9%	79·6%	72·9%
Image plus age plus blastocyst grade	77·6% (75·1–80·1)	72·5%	81·1%	86·4%	73·3%	69·1%	78·9%	80·3%	76·1%
Image plus age plus blastocyst score	77·4% (74·8–79·8)	70·0%	83·3%	87·6%	74·7%	67·9%	76·5%	79·6%	76·1%
Image plus age plus AIBS	75·9% (73·2–78·3)	71·3%	78·7%	87·0%	73·8%	64·2%	75·5%	82·8%	71·9%
Image plus age plus blastocyst grade plus morphokinetics	77·0% (74·4–79·5)	71·3%	81·0%	87·0%	73·3%	67·5%	77·2%	79·6%	76·1%
Image plus age plus blastocyst score plus morphokinetics	77·6% (75·0–80·0)	76·7%	78·0%	87·6%	73·8%	69·4%	76·5%	81·3%	75·4%
Image plus age plus AIBS plus morphokinetics	75·0% (72·4–77·6)	69·4%	78·9%	87·3%	68·0%	64·2%	76·5%	82·3%	70·9%

Data are % (95% CI) or %. A modified ResNet18 architecture was trained, validated, and tested for three classification tasks: aneuploids versus euploids; complex aneuploids versus single aneuploids plus euploids; and complex aneuploids versus euploids. Combinations of the clinical features and images were used to develop the models. Performance measurements include accuracy (95% CI), PPV, and NPV. The test dataset and its predictions were then stratified into demographics on the basis of maternal age and day of blastocyst formation. AIBS=artificial intelligence blastocyst scores. NPV=negative predictive value. PPV=positive predictive value.

Table 3:

Fetal heart and livebirth outcomes by maternal age

	Count	Fetal heart outcomes	Livebirth outcomes
All maternal ages			
Global	242	137 (57%)	118 (49%)
Predicted euploid	166	93 (56%)	80 (48%)
Predicted aneuploid	76	44 (58%)	38 (50%)
Maternal age <35 years			
Global	86	51 (59%)	44 (51%)
Predicted euploid	83	51 (61%)	44 (53%)
Predicted aneuploid	3	0	0
Maternal age ≥35 to <37 years			
Global	67	33 (49%)	28 (42%)
Predicted euploid	52	25 (48%)	21 (40%)
Predicted aneuploid	15	8 (53%)	7 (47%)
Maternal age ≥37 to <39 years			
Global	56	33 (59%)	31 (55%)
Predicted euploid	29	15 (52%)	14 (48%)
Predicted aneuploid	27	18 (67%)	17 (63%)
Maternal age ≥39 years			
Global	33	20 (61%)	15 (45%)
Predicted euploid	2	2 (100%)	1 (50%)
Predicted aneuploid	31	18 (58%)	14 (45%)

Data are n or n (%). 242 transferred embryos labelled as euploid by preimplantation genetic testing for aneuploidy with known fetal heart and livebirth outcomes (total euploid plus aneuploid) were compared with embryos predicted as euploid and aneuploid by STORK-A.