RESEARCH ARTICLE

Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences

Jørgen Berntsen[1]*, **Jens Rimestad**[1], **Jacob Theilgaard Lassen**[1], **Dang Tran**[2], **Mikkel Fly Kragh**[1,3]

1 Vitrolife A/S, Aarhus, Denmark, **2** Harrison AI, Sydney, New South Wales, Australia, **3** Department of Electrical and Computer Engineering, Aarhus University, Aarhus, Denmark

* jberntsen@vitrolife.com

OPEN ACCESS

Citation: Berntsen J, Rimestad J, Lassen JT, Tran D, Kragh MF (2022) Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences. PLoS ONE 17(2): e0262661. https://doi.org/10.1371/journal.pone.0262661

Editor: Marcelo Fábio Gouveia Nogueira, School of Sciences and Languages, Sao Paulo State University (UNESP), BRAZIL

Received: February 25, 2021

Accepted: January 3, 2022

Published: February 2, 2022

Data Availability Statement: The data underlying this study are not publicly available. The data sets are owned by the 18 participating clinics. In the data agreement contract between the clinic and Vitrolife it is explicitly stated that data must not be made public due to potentially sensitive information. The dataset contains anonymized patient demographics, basic treatment information (e.g. start month, insemination method, length of incubation), transfer decisions/outcomes and time-lapse images for each embryo. Data access

Abstract

Assessing and selecting the most viable embryos for transfer is an essential part of in vitro fertilization (IVF). In recent years, several approaches have been made to improve and automate the procedure using artificial intelligence (AI) and deep learning. Based on images of embryos with known implantation data (KID), AI models have been trained to automatically score embryos related to their chance of achieving a successful implantation. However, as of now, only limited research has been conducted to evaluate how embryo selection models generalize to new clinics and how they perform in subgroup analyses across various conditions. In this paper, we investigate how a deep learning-based embryo selection model using only time-lapse image sequences performs across different patient ages and clinical conditions, and how it correlates with traditional morphokinetic parameters. The model was trained and evaluated based on a large dataset from 18 IVF centers consisting of 115,832 embryos, of which 14,644 embryos were transferred KID embryos. In an independent test set, the AI model sorted KID embryos with an area under the curve (AUC) of a receiver operating characteristic curve of 0.67 and all embryos with an AUC of 0.95. A clinic hold-out test showed that the model generalized to new clinics with an AUC range of 0.60–0.75 for KID embryos. Across different subgroups of age, insemination method, incubation time, and transfer protocol, the AUC ranged between 0.63 and 0.69. Furthermore, model predictions correlated positively with blastocyst grading and negatively with direct cleavages. The fully automated iDAScore v1.0 model was shown to perform at least as good as a state-of-the-art manual embryo selection model. Moreover, full automatization of embryo scoring implies fewer manual evaluations and eliminates biases due to inter- and intraobserver variation.

Introduction

Embryo assessment to predict the most viable embryo for transfer has been a challenge since the start of in vitro fertilization (IVF). The introduction of time-lapse in clinical routines [1]

requests can be sent to the individual IVF clinics or clinic chain, contact information listed below: a. Virtus Health Head Office, Level 3, 176 Pacific Highway, Greenwich NSW 2065, Australia. Email: info@ivf.com.au b. Ciconia Fertilitetsklinik, Saralyst Allé 50, 8270 Højbjerg, Denmark. E-mail: aarhus@ciconia.dk c. Skive Fertilitetsklinik, Resenvej 25, 7800 Skive, Denmark. E-mail: fertilitetsklinikken@midt.rm.dk d. Horsens Fertilitsklinik, Sundvej 30, 8700 Horsens, Denmark. E-mail: fertilitet@horsens.rm.dk e. Kato Ladies Clinic, Nishishinjuku, 7, Shinjuku City, 160-0023 Tokyo, Japan. E-mail: klc@towako-kato.com f. Maigaard Fertilitetsklinik, Jens Baggesens Vej 88H, 8200 Aarhus, Denmark. E-mail: mail@maigaard.dk The authors did not receive any special privileges in accessing the data.

Funding: Vitrolife provided support in the form of salaries for J.B., J.R., J.T.L, M.F.K., but did not have any additional role in the study design, decision to publish, or analysis and preparation of the manuscript. Vitrolife supported data collection and equipment for AI model training. Harrison.AI supported data collection. Data collection were done in collaboration with local Vitrolife offices. However, study design, analysis and the decision to publish was made solely by the authors.

Competing interests: This study was supported by Vitrolife, the employer of J.B., J.R., J.T.L, M.F.K. All authors participated in the study design, data collection and analysis and preparation of the manuscript. The decision to publish was taken by J.B. Vitrolife produces and markets iDAScore. The study was also supported by Harrison.AI, the employer of D.T. He participated in the study design, data collection and analysis. D.T. has a patent related the current study. J.B. and J.R. are Vitrolife A/B shareholders. This does not alter our adherence to PLOS ONE policies on sharing data and materials, as detailed online in our guide for authors.

has enabled continuous monitoring of embryo development in vitro. This has allowed for the detection of morphological changes and events with the exact time-point of occurrence [2]. Based on these morphokinetic parameters, several embryo selection models have been developed [3–7]. The end-points of these models were both blastocyst prediction [8, 9], genetic status [10–12], gestational sacs [4] and live birth [13–15]. Studies have shown a general improvement when time-lapse and morphokinetic selection are used compared with standard incubation [16, 17] while other studies have found a need for in-house validation of models before use [18].

During the past decade, a drastic improvement in different technologies used in the IVF labs has been observed, which has implied an increase in workload [19]. Even though morphokinetic analysis facilitates a flexible workflow, improved assessment consistency and reduce inter- and intra-observer variations [20–23], there is still considerable challenges to manage embryo selection in a busy lab. This has led to the development of automatic scoring of morphology and morphokinetics. Initial approaches were based on traditional computer vision technology [24, 25], while recent methods have used more powerful deep learning frameworks [26–31]. The outputs from these methods are used as inputs for decision support systems for ranking or selecting embryos for transfer or cryopreservation.

To completely skip the intermediate steps of estimation of morphology and/or morphokinetics, several newly published methods directly estimate the outcome of an embryo transfer. This can be done based on single images using traditional computer vision technologies [32] or more advanced deep learning technologies [33, 34]. Covering the whole dynamic embryo development, time-lapse sequences have been used to predict pregnancy in terms of fetal heartbeat (FH) using deep learning [35]. This fully automated method obtained an area under the curve (AUC) of the receiver operating characteristic (ROC) of 0.93 for sorting the whole cohort of available embryos in relation to FH.

As deep learning models are typically overparameterized, they easily overfit to the training data set. Therefore, emphasis must be placed on obtaining and evaluating generalization ability across different clinical practices in order to ensure a reliable and trustworthy model that is safe to use in new clinics. To ensure this, a model must be developed and evaluated using very large data sets that cover a range of different clinical practices, incubation conditions, insemination methods and patient cohorts.

In general, deep learning methods can be considered as "black boxes" as interpretation is not transparent. Thus, it has been discussed if biological justification is required for acceptance of computer-generated algorithms to select embryos based on machine-learned combinations of parameters [36]. Based on the above, the current analysis aims to develop a fully automated embryo selection model based on a deep learning network that is generalizable, robust and can be supported by biological evidence.

Materials and methods

Description of data

Retrospective data from 18 clinics worldwide from between 2011 and 2019 were included in the investigation. No specific methods were applied in the selection of the clinics. Table 1 lists the total number of embryos, the number of transferred embryos and the average female age for each clinic. All data used in the studies were retrospective and provided in a de-identified format. In Denmark, the study described was deemed exempt from notification to the National Committee on Health Research Ethics according to the Act on Research Ethics Review of Health Research Projects (consolidation act no. 1338 of September 1, 2020).

Table 1. Distribution of total number of embryos, fresh embryos and thawed embryos with known implantation data (KID), total number of annotated embryos and average female age.

Clinic	Total number of embryos	Fresh KID embryos	Thawed KID embryos	Annotated embryos	Mean age
1	24360	960	21	14139	32.0
2	18990	1182	1490	4124	37.2
3	13165	550	777	8810	36.9
4	12640	536	166	3875	32.8
5	10138	833	-	5769	36.6
6	7679	65	2768	6498	41.3
7	5773	386	509	24	37.6
8	5757	526	451	2588	36.4
9	4215	373	242	0	36.4
10	3316	455	429	2839	37.2
11	2729	406	185	1156	34.8
12	1864	228	152	767	35.6
13	1098	140	52	486	31.8
14	1042	76	114	337	36.2
15	1007	89	113	330	-
16	817	100	69	775	35.9
17	759	72	54	0	36.0
18	483	71	4	376	36.6
total	115832	7048	7596	52893	36.0

https://doi.org/10.1371/journal.pone.0262661.t001

All embryos were cultured in the EmbryoScope-D™ or EmbryoScope Plus™ incubators (Vitrolife A/S, Aarhus, Denmark). Image sequences were acquired using the settings in each clinic. The number of focal planes varied from 3 to 11. The average time between image acquisition was 15 and 11 minutes for embryos incubated in EmbryoScope-D™ and EmbryoScope Plus™, respectively.

The average patient age was 35.6 with a range from 18 to 52 years. No exclusion of patients was performed. All embryos inserted into the EmbryoScope incubators before 24 hours post insemination (hpi) and removed after 108 hpi were included. Thus, fresh oocytes, cryopreserved oocytes and donated oocytes were all included. Embryos that underwent assisted hatching and biopsy for preimplantation genetic testing (PGT) were also included. On average, each cycle contained 5.7 embryos with a standard deviation of 4.6.

Embryos were classified as FH+, FH-, Discard, Unknown or Pending similar to the procedure used for the IVY model [35]:

- FH+ is when the number of FHs observed on ultrasound after 7 weeks was equal to the number of embryos transferred.
- FH- is when no pregnancy occurred or no FH was observed on ultrasound.
- Discarded embryos were manually deselected by the embryologist. In the final model training, discarded embryos were pseudo-labelled as FH-.
- Unknown is when the number of transferred embryos was greater than the number of FHs observed on ultrasound.
- Pending is when the embryo was cryopreserved and no outcome was recorded yet.

In total, 14,644 embryos were transferred (fresh or cryopreserved), resulting in 4,337 FH + and 10,307 FH- outcomes. In addition, 101,188 embryos were labelled as discarded due to manual deselection by embryologists or due to aneuploidy (1,610 embryos). 23,002 embryos were classified as pending, and 754 embryos were classified as unknown and were not included in the data set. The final data set consisted of 115,832 embryos labelled as either FH+, FH- or discarded. The data set was randomly split into a training data set (85%) and an independent test data set (15%). This was done for individual embryos across all treatments and clinics.

Deep learning model

An AI model was trained for binary classification of the positive and negative FHs for the 98,583 embryos in the training data set. Python 3.6.5 and TensorFlow v2.0 were used for the model development. Training was performed using two Nvidia Quadro RTX8000 GPUs.

The overall deep learning network structure for prediction of FH is shown in Fig 1. The inflated 3D (I3D) convolutional neural network [37], a state-of-the-art network for video action recognition, was the basis of the network. The default width of the I3D network was reduced to 25% everywhere. The output from the I3D network was pooled using spatial maximum and average pooling and then concatenated. The output was fed to a bidirectional LSTM [38] with 128 units in each direction resulting in 256 outputs. These were fed to two separate fully connected layers for binary classification of FH and the auxiliary output discard, both with sigmoid activation functions. The model was trained using the Adam optimizer [39] and the one-cycle-policy [40] with an initial learning rate of 1e-5 and a maximum learning rate of 1e-4. Furthermore, a batch size of 64 and a dropout of 25% were used. Hyperparameter estimation and initial experiments were performed with 5-fold cross-validation on the training data. The final model was trained on the whole training data set without a validation data set [41, 42]. The final training used the hyperparameter settings found during the 5-fold cross-validation.

The input to the model was 128 frames sampled one hour apart with one focal plane and a resolution of 256x256 pixels. The sequence was offset by 12 hpi, thus effectively covering 12–140 hpi.

An important hyperparameter is the method used for sampling of embryos during training. This was investigated in a 5-fold cross-validation experiment by varying the relative distribution between FH+ KID embryos, FH- KID embryos and discarded embryos. In the first two experiments, only transferred KID embryos were included, either unbalanced according to the prevalence of 30% FH+ / 70% FH-, or balanced with a distribution of 50% FH+ / 50% FH-. In the last experiment, discarded embryos were also included, resulting in a distribution of 50% FH+ / 10% FH- / 40% discarded embryos. In this experiment, the discarded embryos were pseudo-labelled as FH-.

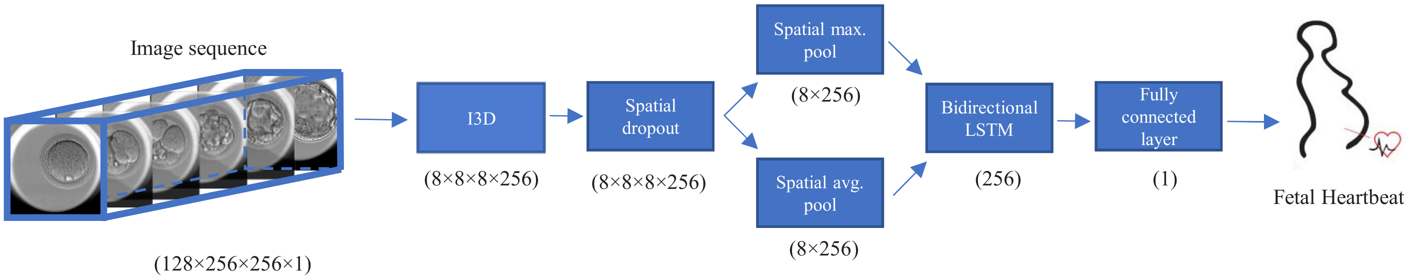

Fig 1. Overall architecture of the deep learning network used for prediction of fetal heartbeat. The image sequence is fed to a 3D CNN based on the I3D architecture. The output from the CNN is reduced in dimensions and fed to an LSTM network to utilize temporal information. Finally, a fully-connected layer predicts the fetal heartbeat. The numbers in the figure show the output dimensions from each step in the sequence.

https://doi.org/10.1371/journal.pone.0262661.g001

Input data were augmented in two stages. First, the whole sequence was offset by a random time in [-0.5, 0.5] hours, the focal plane was sampled at –1, 0 and +1 from the central focal plane with a probability of 25%, 50% and 25%, respectively, jitter was introduced by random time shifts of +/- 0.375 hours, and the end time of the sequence was randomly truncated to between 108 and 140 hpi. The sequence was finally padded with zeros to always include 128 frames. Next, the sampled data sequence was applied temporal coherent random cut out [43], 90–110% rescaling, up to 10% translation horizontally and vertically, brightness, horizontal flipping and up to 10-degree rotation clockwise and counter clockwise. Training was done for 9,264 mini batches, corresponding to 80 epochs for FH+ embryos. The model predicted both FH and discard. For both outputs, the focal loss function [44] with a gamma of 2.0 and an alpha of 0.5 was used. Finally, after training, the FH output from the model was rescaled from [0, 1] to [1.0, 9.9].

Clinic hold-out test

To evaluate how the model generalizes to new clinics, a leave-one-clinic-out cross-validation was performed on the 12 clinics that each had more than 250 KID embryos. That is, 12 unique models were trained on the remaining clinics (17 clinics in total). The training was performed using the same settings. The trained model was then evaluated on all data from the test clinic.

Morphokinetic annotations

In some of the clinics, embryo morphology and morphokinetic events were recorded (Table 1). Annotations were performed according to guidelines by Vitrolife [45] and published nomenclature [2]. The following annotations were used: number of pronuclei (PN), cleavage to 2-blastomeres (t2), 3-blastomeres (t3) and 5-blastomeres (t5), time to full blastocyst (tB) and grading of inner cell mass (ICM) and trophectoderm (TE).

Direct cleavage is a term describing the phenomenon in which a blastomere divides into three or more cells during a single cell cycle. With time-lapse imaging, direct cleavages have been measured using morphokinetic annotations studies [46]. Here, we define a direct 1 to 3 division (DC1-3) when t3-tPNf < 5h and a direct 2 to 5 division (DC2-5) when t5-t3 < 5h. If DC1-3 or DC2-5 was not observed, the embryo was defined as having no direct cleavages (no-DC).

For cycles with enough morphokinetic annotations, the KIDScore D5 v3 [47] score was calculated. The model requires annotations of PNs, t2, t3, t5, tB, ICM and TE.

Statistical analysis

R software version 3.5.3 was used for statistical analysis. For the evaluation of classification performance, the pROC-package [48] was used for estimating ROC curves and calculating the AUC. Tests for significant differences between AUCs were carried out using either paired or unpaired DeLong's one-tailed test. The Mann–Whitney–Wilcoxon test was used to test differences in morphokinetic variables between different groups. A paired student t-test was used to test for significant differences between the performances of two models. A $p < 0.05$ was considered significant.

Results

Initial experiments

The effect of how embryos are sampled from the training data set was tested using three 5-fold cross validation experiments. In the first two experiments, the model was trained exclusively on transferred KID embryos which were either sampled according to actual prevalence (30% FH+ and 70% FH-) or with oversampling of positive samples (50% FH+ and 50% FH-). In the third experiment, the model was trained with discarded embryos pseudo-labelled as FH- and with oversampling of positive samples (50% FH+, 10% FH- and 40% discarded). The performance was evaluated on transferred KID embryos in the validation data set for each of the 5 folds. There were no significant differences ($p > 0.05$) in mean AUC between the three different sampling strategies (Table 2).

The overall performance on all embryos was also evaluated. The sampling strategy where discarded embryos were included had a significantly higher mean AUC ($p < 0.005$) compared to the two other sampling strategies. There was no significant difference in the overall performance between the two strategies where discarded embryos were not included. The score distribution of the FH+, FH- and discarded embryos showed different patterns for the three strategies (S1 Fig). For models only trained on KID embryos, the score distribution of discarded embryos overlapped with scores of transferred FH- and FH+ embryos. In contrast, for models trained on both KID embryos and discarded embryos there was less overlap between discarded embryos and the transferred embryos. Based on these cross-validation experiments, all subsequent experiments were performed with both KID embryos and discarded embryos with a sampling of 50% FH+, 10% FH- and 40% discarded embryos.

Final model

All subsequent analyses were performed on the independent test data set that was not used during training. Totally, this data set contained 17,249 embryos of which 2,212 were transferred embryos with known outcome (KID embryos). The overall model performance was evaluated for its sorting capability using AUC. For KID embryos, the AUC was 0.67 with a 95% confidence interval of 0.64–0.69 (Table 3). If the whole cohort was considered, the AUC was 0.95 with a 95% confidence interval of 0.95–0.96.

Sub-group analysis

The following sub-groups were analysed: patient age (for the groups <30, 30–34, 35–39 and >39 years), insemination method (IVF or ICSI), length of incubation (5 or 6 days) and fresh vs cryopreserved embryo transfer (Table 3). For age groups, the AUC for KID embryos varied between 0.63 and 0.69, with the lowest AUC for the age group 30–34 years. With regards to insemination method, the AUC was 0.69 and 0.67 for ICSI and IVF, respectively. For length of incubation, the AUC was 0.65 and 0.66 for D5 and D6, respectively. Cryopreserved embryos had a significantly lower AUC of 0.65 compared to fresh transfers with an AUC of 0.69.

Clinic hold-out test

To investigate how the chosen model architecture and training data generalize to new clinics, a clinic hold-out test was performed. The AUCs for the individual clinics varied between 0.60

Table 2. Five-fold cross-validation experiment on varying the distribution between FH+, FH- and discarded embryos used during training. The area under the curve (AUC) was evaluated for both transferred embryos with known implantation data (KID) and all embryos for each validation data set.

Distribution			Validation AUC	
FH+	FH-	Discarded	KID embryos (mean ± st. dev)	All embryos (mean ± st. dev)
30%	70%	0%	0.680 ± 0.016	0.903 ± 0.018
50%	50%	0%	0.686 ± 0.007	0.907 ± 0.007
50%	10%	40%	0.687 ± 0.011	0.954 ± 0.002

https://doi.org/10.1371/journal.pone.0262661.t002

Table 3. Number of transferred embryos with known implantation data (KID), area under the curve (AUC) and 95% confidence intervals (C.I.) for the following sub-groups: age, insemination method, length of incubation and fresh vs cryopreserved embryo transfer.

Parameters	Sub-group	Number of KID embryos	AUC	95% C.I.
Overall		2212	0.67	0.64–0.69
Age	<30	177	0.69	0.61–0.77
	30–34	444	0.63	0.58–0.68
	35–39	655	0.67	0.63–0.72
	>39	598	0.66	0.61–0.72
Insemination method	ICSI	738	0.69	0.65–0.73
	IVF	343	0.67	0.62–0.73
Length of incubation	D5	1894	0.65	0.63–0.68
	D6	303	0.66	0.57–0.74
Transfer protocol	Fresh	1070	0.69	0.66–0.72
	Cryopreserved	1142	0.65*	0.61–0.68

A significantly lower AUC for a sub-group compared to the AUC for the remaining sub-groups is indicated with a star ($p < 0.05$). In addition, a precision-recall (PR) curve was calculated for the KID embryos (S2 Fig).

https://doi.org/10.1371/journal.pone.0262661.t003

Table 4. The area under the curve (AUC) and 95% Confidence Interval (C.I.) in the clinic hold-out test for embryos with known outcome (KID).

Clinic	AUC	95% C.I.
1	0.63*	0.59–0.66
2	0.66	0.64–0.68
3	0.65	0.61–0.68
4	0.60*	0.56–0.65
5	0.75	0.71–0.78
6	0.72	0.70–0.74
7	0.67	0.64–0.72
8	0.66	0.62–0.70
9	0.65	0.60–0.70
10	0.65	0.61–0.69
11	0.68	0.64–0.73
12	0.68	0.62–0.73

Only clinics with more than 250 KID embryos were included. These clinics are identical with the first 12 clinics in Table 1. A significantly lower AUC for a clinic compared to the AUC for the hold-out data set is indicated with a star ($p < 0.05$).

https://doi.org/10.1371/journal.pone.0262661.t004

and 0.75 (Table 4). For clinic 1 and 4, the AUC was significantly smaller than for the rest of the clinics included in the test data.

Correlation with morphology and morphokinetic annotations

The biological validity of the model was evaluated by correlating the scores with morphological and morphokinetic parameters. This comparison was only done for the embryos in the test data set that had morphological and morphokinetic annotations.

Responses to direct cleavages were tested on the subset of the test data for which the timings t2, t3 and t5 were all annotated (Fig 2A). For the whole embryo cohort, the scores were significantly different across all three groups ($p < 0.0001$). For the blastocyst group (i.e. with a tB annotation), the scores in the no-DC group were significantly different from both DC1-3 and

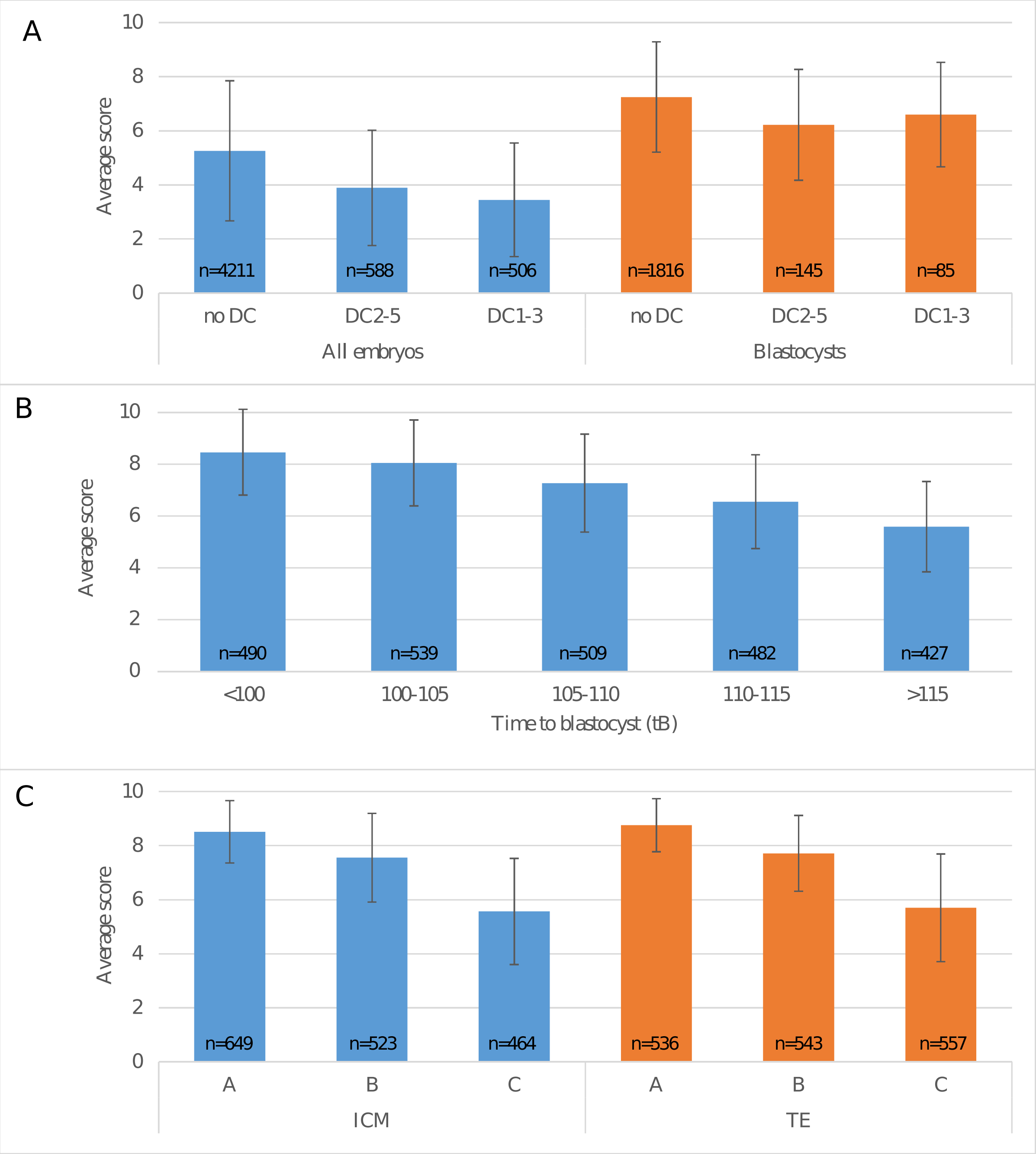

Fig 2. iDAScore for different morphokinetic groups. (A) Average score for embryos with no direct cleavages (no DC), direct cleavage from 2 to 5 blastomeres (DC2-5) or direct cleavage from 1 to 3 blastomeres (DC1-3), respectively. Average scores are shown for the whole cohort (n = 5305) and for embryos that reached the blastocyst stage (n = 2046). (B) Average score for blastocysts (n = 2447) where time to blastocyst was <100, 100–105, 105–110, 110–115 or >115 hpi, respectively. (C) Average score for embryos (n = 1636) with ICM and TE grades A, B or C, respectively. The number of embryos in each group is shown inside the bar. Lines show the standard deviation within each group.

https://doi.org/10.1371/journal.pone.0262661.g002

DC2-5 ($p < 0.0001$). However, there was no significant difference between DC1-3 and DC2-5 ($p = 0.18$).

Responses to development speed were tested on the subset of the test data for which tB was annotated (Fig 2B). The average score decreased with slower development and there were significant differences between the average scores in all tB groups ($p < 0.0001$). For embryos with annotated ICM and TE, the average score decreased with lower grades (Fig 2C) and there were significant differences between different grades ($p < 0.0001$).

Comparison with the KIDScore D5 v3 model

The performance of the model was compared with the KIDScore D5 v3 model [47] in cycles where enough morphological and morphokinetic parameters were annotated. For KID embryos, only embryos with annotation of tB, ICM and TE were included. Of the total 17,249 embryos in the test set, there were 7,932 annotated embryos, and of the 2,212 KID embryos, there were 1,094 annotated embryos. The ROC curve was calculated for both models and for both the whole cohort and the KID embryos (Fig 3). For sorting the whole cohort, the AUC of 0.92 for the iDAScore v1.0 model was significantly higher than the AUC of 0.89 for KIDScore D5 v3. Note that the AUC of 0.92 is different from the above whole cohort-AUC of 0.95 as the evaluation was only done using the subset of embryos with manual annotations (Table 1). For sorting of KID embryos, there was no significant differences between the AUC of 0.67 for the iDAScore v1.0 model and 0.66 for KIDScore D5 v3.

Discussion

Model training

A fundamental problem when training embryo selection models based on implantation data is that only a fraction of the whole cohort is labelled with a known outcome. The remaining embryos are not labelled as they have no known outcome. In addition, the labelled embryos are often unbalanced as negative labels are more predominant than positive labels. These features of the training data and the methods used for training might give rise to potential biases [49]. We have investigated these biases by testing three different sampling strategies for handling missing labels and unbalanced data: training on KID embryos with no oversampling, training on KID embryos with oversampling of FH+ and training on both KID embryos and pseudo-labelled discarded embryos. The results showed that the inclusion of discarded embryos did not have any effect on the performance when evaluated on KID embryos. However, when evaluated on all embryos, it was clear that the performance with regards to recognizing poor embryos significantly increased (Table 2 and S1 Fig). Thus, in general the inclusion of discarded embryos in a model training will facilitate a fully automated evaluation applicable for all embryos in a cohort. In contrast, when models are trained only on KID embryos, the user must perform a pre-selection step where transferrable embryos must be identified before using such a model.

Final model

The sorting capability for the whole cohort had an AUC of 0.95, which is higher than the 0.93 observed with the IVY model [35]. It should be noted that all clinics in the IVY model also contributed with data in this investigation. However, additional new data from these clinics and 5 new clinics were added in this investigation so that the total number of embryos increased from 8,836 to 115,832, the number of embryos with known implantation data (KID) increased from 1,773 to 14,644 and the number of FH+ embryos increased from 694 to 4,337.

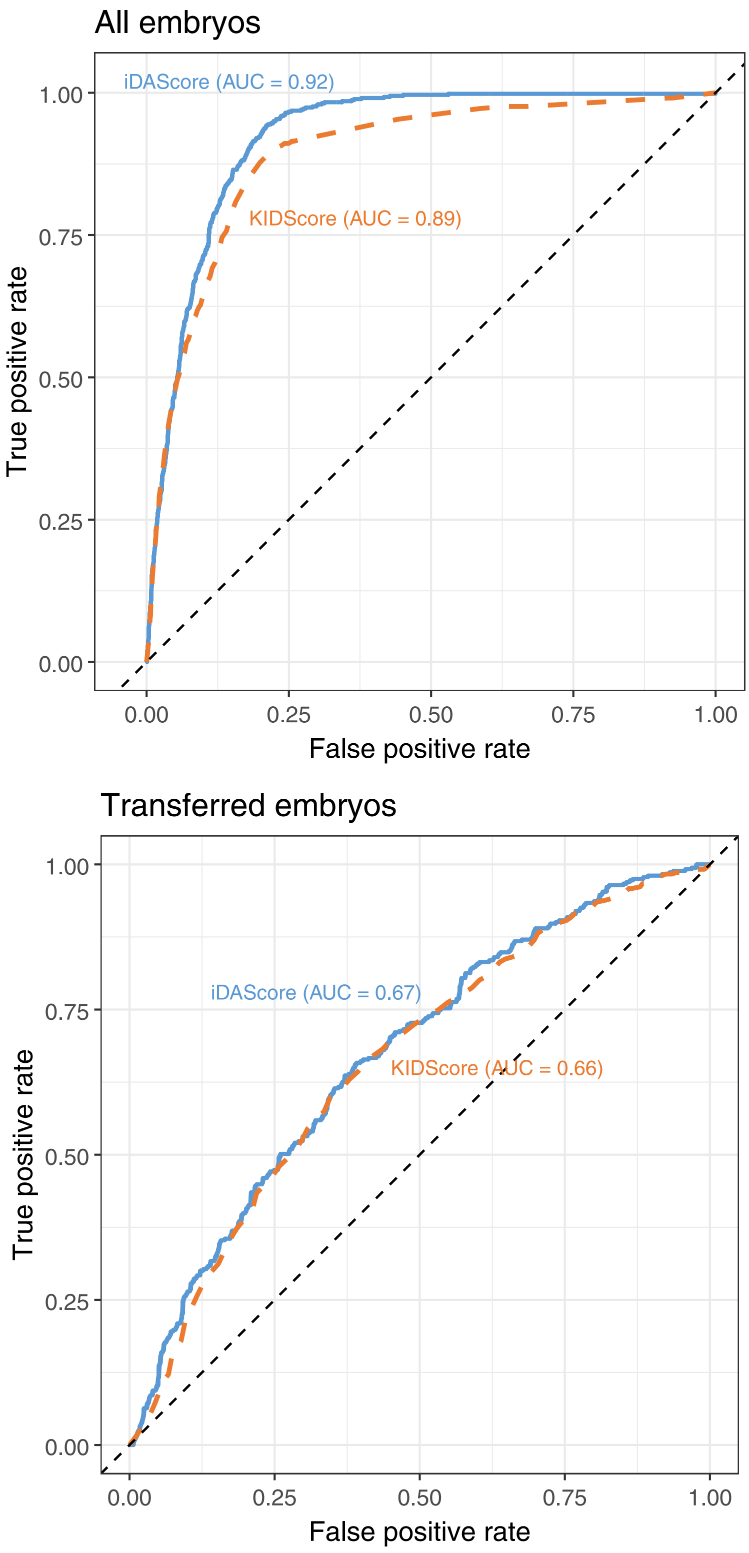

Fig 3. Comparison between ROC curves based on iDAScore (solid blue) and KIDScore (dotted orange). The upper plot shows the ROC curve for the whole embryo cohort (n = 7,932), and the lower plot shows the ROC curve for KID embryos (n = 1,094).

https://doi.org/10.1371/journal.pone.0262661.g003

Thus, the current data set was more than 6 times larger than the data set used for training the IVY model, which means that an even higher performance and robustness is obtained. To our knowledge, this is currently the largest data set used for developing an embryo selection model.

The model's sorting capabilities for KID embryos had an AUC of 0.67 (Table 3) which is lower than the sorting capability for the whole cohort, where the AUC was 0.95. This difference is expected, as it is much more difficult to sort between good morphology blastocysts than to sort within the whole cohort that includes a range from arrested embryos to good morphology blastocysts. Which measure is the most relevant depends on the task of the model. If the purpose is to have a model that is fully automated and sorts between all embryos, the high AUC for sorting the whole cohort is the most relevant measure. However, if the user prescreens the embryos to find potential transfer candidates, the lower AUC for sorting only within the KID embryos is the most relevant. As both tasks are relevant for embryo selection algorithms and as practical use will depend on the actual clinical setting, we propose that both measurements should be reported in future publications.

Sub-group analysis

When designing a selection model that is intended to be used across a wide range of clinical settings, patients and culture conditions, it is extremely important to investigate any potential bias within sub-groups in the data set. For nearly all sub-groups, the AUC was within the 95% confidence interval of the AUC of the overall ROC curve. Only the cryopreserved sub-group had a significantly lower AUC than the other sub-groups. The reason for this is probably that other processes (e.g. the vitrification/warming and subsequent endometrium preparation) have an impact on a successful implantation. Thus, the sorting becomes more difficult, which results in a lower AUC.

To our knowledge, no other selection models have been tested on different sub-groups within a large independent data set. However, it is very important to check that models perform comparable on different sub-sets and do not exhibit general biases. When a selection model uses age as a parameter, testing on age sub-groups becomes even more important. The inclusion of age will inevitably improve the overall model performance significantly, as age is one of the best predictors for successful implantation. However, including age may not improve the sorting capability on treatment level (i.e. the embryo cohort of a single patient), which ultimately is the essence of an embryo selection model. Both the STORK algorithm [28] and AIR E [32] include age as input in their models. However, as no age sub-group analysis was performed, the sorting capabilities of these models on treatment level remain to be documented.

The sub-group analysis is an internal validation where the model is tested on a fraction of the training data [50]. Potentially this presents an optimistic performance estimate as the test data has the same patient population and distribution as in the training data. This issue can be addressed using a geographic internal validation (a clinic hold-out test) or even stronger by using an external validation, where the model is tested under new conditions such as new clinics, time-periods, procedures, or populations that were not used during model development [50].

Clinic hold-out test

Several studies have shown the need for testing selection models on in-house data before clinical use [18, 51]. In this analysis, we used a clinical hold-out validation approach where the model was first trained on data excluding a specific clinic and then tested on this clinic. As shown in Table 4, these models showed a similar sorting capability with the majority being within the 95% confidence interval of the final model. Clinic 1 and 4 with young women (Table 1) had a significantly lower AUC than on the test set. For the younger women in these clinics, the majority of the transferred KID embryos were probably top-quality blastocysts. For the older women in the other clinics, it is likely that a more diverse cohort of KID embryos are transferred. As it is more difficult to sort within a group of homogeneous top-quality blastocysts than within a more diverse group of KID embryos, the AUCs were lower for clinics 1 and 4. Thus, the difference in AUC is probably not related to the performance of the model but rather to a bias in evaluating on pre-selected embryos for transfer.

It should be noted that a high AUC does not imply a high implantation rate or vice versa. A high AUC specifically refers to how a model sorts embryos within a given cohort. Thus, differences in AUCs can result from poor model generalization, but indeed also from differences in patient cohort and clinical practices as above.

Correlation with morphology and morphokinetic annotations

For direct cleavages, Fig 2A shows a clear decrease in the average score for the whole cohort of embryos with direct cleavages annotated. This agrees with several day 3 embryo grading models [4, 5] where embryos with direct cleavages are excluded or given a low score. It should also be noted that DC1-3 embryos had a lower score than DC2-5 embryos which agrees with other studies [46] on day 3 transfers. Furthermore, it was observed that for blastocysts, the impact of direct cleavages on the score was smaller than for the whole cohort. This aligns with the hypothesis of "self-correction" where the embryo can correct mitotic cell divisions errors and, if the blastocyst stage is reached, have a relatively better implantation rate compared to a day 3 transfer [46, 52].

A clear significant effect was observed between the score and time to blastocyst. This has also been investigated in several other studies [36]. The timing and grade of blastocyst expansion have been shown by several investigators to be important predictors of implantation [53, 54]. In addition, time to blastocyst is also a direct parameter in the KIDScore D5 logistic regression model [47].

For both ICM and TE, there was a clear correlation between the score and the grade. This supports several studies that have shown how ICM and TE grades correlate with implantation [26, 54–56].

Thus, the above analyses clearly demonstrate that iDAScore outputs correlate with the most established biologically meaningful parameters that are known to have an impact on implantation. It should be underlined that the model was never trained on any of these parameters, but apparently indirectly learned features that correlate with these biologically meaningful parameters.

Selection bias

When comparing AI models there will always be biases due to the used embryo selection practice [50]. Due to the large number of clinics and the wide span of years in the current study, a large range of different embryo selection methods were used. This includes at least the use of both KIDScore D5, Gardner grading, various clinic specific models and PGT. Our data do not include information on which embryo selection model was used for each specific transfer, so it is not possible to perform a general evaluation of selection biases.

In some cases, embryo selection was based on PGT results. In the total data set, there were 654 transferred euploid KID embryos (4.5% of all transfers) and 1,610 discarded aneuploid embryos (1.6% of all discarded). However, as the test data only included 86 euploid KID embryos, it was not possible to perform a sub-group analysis for this group specifically.

Model comparisons

Currently, there is no general agreement on how to compare the performance of different embryo selection models [50]. Some studies have measured model performance using well-known diagnostic performance parameters such as sensitivity, specificity and accuracy [57, 58], while other studies have used AUC to measure the sorting capabilities [4, 28, 35]. Measurements such as sensitivity, specificity and accuracy are excellent for evaluating the overall performance of a diagnostic test. Usually, these models have a binary output: positive or negative. However, in a clinical setting where the task is to rank embryos, we believe that AUC is the best measurement for models that have more than two outputs. The AUC values integrate the sensitivity and specificity over the whole range of possible score values. In that way, an estimate of the overall sorting capability of the model is obtained. In addition, based on a ROC curve (Fig 3) it is possible to estimate the sensitivity and specificity at any score threshold.

The best method for comparing different models is to evaluate AUC on the same independent data set. In the present work, we compared a fully automated model (iDAScore v1.0) based on deep learning with the state-of-the-art KIDScore D5 v3 morphokinetic model based on manual subjective labelling. In this comparison, the deep learning model had a significantly higher AUC for sorting within the whole cohort of embryos (Fig 3). The AUC for KID embryos was numerically higher for the fully automated model, but there was no significant difference. Thus, the fully automated model in general performs better than the more subjective and annotation-based KIDScore D5 v3 model.

Based on the ROC curve for the whole cohort, it was observed that the iDAScore v1.0 and KIDScore D5 v3 perform nearly identical up to a true positive rate of 0.5. This part of the curves is based on the high scoring embryos. Thus, for these embryos, the sorting capability is nearly identical. However, above the true positive rate of 0.5, the ROC curve for iDAScore v1.0 becomes higher than for KIDScore D5 v3. This part of the curve is based on embryos with intermediate scores where iDAScore v1.0 performs significantly better than KIDScore D5 v3. The reason for this is that iDAScore v1.0 was also trained on discarded embryos, while KIDScore D5 v3 was only trained on transferred KID embryos. Thus, if the aim of a model is a fully automated system with the purpose of sorting all embryos in a cycle, it is important that the model learns both how transferred embryos implant and how lower scoring embryos perform. In example, if arrested cleavage stage embryos are not part of the training data, the model's performance on this kind of embryos will be unpredictable as the model has never seen this type of embryos before (see S1 Fig). This can be obtained by including discarded embryos as in the current investigation.

The current study describes the first version of the iDAScore model. Future versions of the model will be improved by including even more training data and using more advanced training and network architectures. In example, this could include better use of transfer learning or pre-training on unlabelled data. Promising methods for pre-training of models include temporal cycle consistency [59] and contrastive learning [60]. Other potential improvements might be the use of new model architectures like video transformers [61, 62] that can replace the LSTM layer or even the 3D convolutions. In addition, the model might also be improved by including other data sources like patient demographics, diagnosis, PGT-A results, proteomic profiles, fluorescent-lifetime imaging microscopy and other types of imaging.

Conclusion

In summary, we have developed a fully automated deep learning model (iDAScore v1.0) based on 115,832 embryo time-lapse sequences. The selection model has an AUC of 0.67 for KID embryos. It was demonstrated that this fully automated model performs better than the state-of-the-art morphokinetic KIDScore D5 v3 model. The high performance was obtained without the need for assessment/annotation by the embryologist.

In addition, it was demonstrated that the model was applicable to clinics excluded from the training data. When the model was tested on different stratifications within the test data set, it was also shown that the model can be generalized across age groups, insemination methods and length of incubation. The sorting capability within the different stratifications was close to the overall performance of the final model on the test data set.

Finally, iDAScore v1.0 score correlated with well-established embryo development and morphology parameters. Thus, the model gave lower scores to embryos that had direct cleavages, high scores to faster developing blastocysts and high scores to embryos with high-quality trophectoderm and inner cell mass. Even without training on these parameters, the model has thus indirectly learned that these parameters are important for implantation.

Supporting information

S1 Fig. Score distribution for different data sampling strategies. Density plots of the FH+ (green), FH- (blue) and discarded (red) embryos evaluated on a single validation fold for the three different sampling strategies. The left panel shows the distribution of the raw sigmoid output from a model that was trained only on KID embryos. In the mid panel a model was trained only on KID embryos but with oversampling of FH+ embryos. In the right panel a model was trained using oversampling of FH+ and including discarded embryos.
(PDF)

S2 Fig. PR curve for KID embryos. Precision/recall curve for KID embryos in the test data set (n = 2,212) based on iDAScore predictions. The dotted line shows the prevalence of FH+ in the data set.
(PDF)

Acknowledgments

We highly appreciate the collaboration and exchange of data from the following clinics: Aagaard Fertility, Ciconia Fertility Clinic, Complete Fertility Centre Southampton, Fertility Clinic Skive, Horsens Fertility Clinic, IVF Australia-Alexandria, IVF Australia-Canberra, IVF Australia-Greenwich, IVF Australia-Hunter, IVF Australia-Westmead, Kato Ladies Clinic, Maigaard Fertility Clinic, Melbourne IVF, Queensland Fertility Group, SIMS Cork, SIMS Dublin, SIMS Rotunda and TasIVF Hobart. The authors would also like to acknowledge Csaba Pribenszky, Mark Larman, Markus Montag, Mia Steen Johnsen and Tine Qvistgaard Kajhøj for their invaluable support and suggestions to the manuscript. Also, thanks to reviewers for their helpful suggestions and comments.

Author Contributions

Data curation: Jørgen Berntsen, Dang Tran.

Formal analysis: Jørgen Berntsen, Jens Rimestad, Jacob Theilgaard Lassen, Dang Tran, Mikkel Fly Kragh.

Investigation: Jens Rimestad, Jacob Theilgaard Lassen.

Writing – original draft: Jørgen Berntsen, Jens Rimestad, Jacob Theilgaard Lassen, Dang Tran, Mikkel Fly Kragh.

Writing – review & editing: Jørgen Berntsen, Jens Rimestad, Jacob Theilgaard Lassen, Mikkel Fly Kragh.

References

1. Pribenszky C, Mátyás S, Kovács P, Losonczi E, Zádori J, Vajta G. Pregnancy achieved by transfer of a single blastocyst selected by time-lapse monitoring. Reproductive BioMedicine Online. 2010; 21 (4):533–6. https://doi.org/10.1016/j.rbmo.2010.04.015 PMID: 20638906
2. Ciray HN, Campbell A, Agerholm IE, Aguilar J, Chamayou S, Esbert M, et al. Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group. Human Reproduction. 2014/10/24. 2014 Dec [cited 2020 Jan 13]; 29(12):2650–60. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25344070
3. Herrero J, Meseguer M. Selection of high potential embryos using time-lapse imaging: the era of morphokinetics. Fertility and sterility. 2013/02/06. 2013 Mar 15 [cited 2020 Jan 13]; 99(4):1030–4. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23395415
4. Petersen BM, Boel M, Montag M, Gardner DK. Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on Day 3. Human Reproduction. 2016/09/08. 2016 Oct 1 [cited 2020 Jan 13]; 31(10):2231–44. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27609980
5. Meseguer M, Herrero J, Tejera A, Hilligsøe KM, Ramsing NB, Remohí J, et al. The use of morphokinetics as a predictor of embryo implantation. Human Reproduction. 2011/08/09. 2011 Oct [cited 2020 Jan 13]; 26(10):2658–71. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21828117
6. Milewski R, Ajduk A. Time-lapse imaging of cleavage divisions in embryo quality assessment. Reproduction. 2017/04/13. 2017 Aug [cited 2020 Jan 13]; 154(2):R37–53. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28408705
7. Ramsing NB, Berntsen J, Callesen H. Automated detection of cell division and movement in time-lapse images of developing bovine embryos can improve selection of viable embryos. Fertility and Sterility. 2007; 88:S38.
8. Conaghan J, Chen AA, Willman SP, Ivani K, Chenette PE, Boostanfar R, et al. Improving embryo selection using a computer-automated time-lapse image analysis test plus day 3 morphology: results from a prospective multicenter trial. Fertility and sterility. 2013/05/28. 2013 Aug [cited 2020 Jan 13]; 100 (2):412–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23721712
9. Motato Y, de los Santos MJ, Escriba MJ, Ruiz BA, Remohí J, Meseguer M. Morphokinetic analysis and embryonic prediction for blastocyst formation through an integrated time-lapse system. Fertility and sterility. 2016 Feb [cited 2020 Jan 13]; 105(2):376–84.e9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26598211
10. 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. Reproductive BioMedicine Online. 2013/02/19. 2013 May [cited 2020 Jan 13]; 26(5):477–85. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23518033
11. Campbell A, Fishel S, Bowman N, Duffy S, Sedler M, Thornton S. Retrospective analysis of outcomes after IVF using an aneuploidy risk model derived from time-lapse imaging without PGS. Reproductive biomedicine online. 2013/05/09. 2013 Aug [cited 2020 Jan 13]; 27(2):140–6. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23683847
12. Minasi MG, Colasante A, Riccio T, Ruberti A, Casciani V, Scarselli F, et al. Correlation between aneuploidy, standard morphology evaluation and morphokinetic development in 1730 biopsied blastocysts: a consecutive case series study. 2016/09/02. 2016 Oct [cited 2020 Jan 13]; 31(10):2245–54. Available from: https://academic.oup.com/humrep/article-lookup/doi/10.1093/humrep/dew183
13. Fishel S, Campbell A, Foad F, Davies L, Best L, Davis N, et al. Evolution of embryo selection for IVF from subjective morphology assessment to objective time-lapse algorithms improves chance of live birth. Reproductive biomedicine online. 2019 Oct 17;S1472-6483(19)30756-4. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31831370
14. Rienzi L, Cimadomo D, Delgado A, Minasi MG, Fabozzi G, del Gallego R, et al. Time of morulation and trophectoderm quality are predictors of a live birth after euploid blastocyst transfer: a multicenter study. Fertility and sterility. 2019 Dec; 112(6):1080–93. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31843084

15. Reignier A, Girard J-M, Lammers J, Chtourou S, Lefebvre T, Barriere P, et al. Performance of Day 5 KIDScore™ morphokinetic prediction models of implantation and live birth after single blastocyst transfer. Journal of Assisted Reproduction and Genetics. 2019/08/23. 2019; 36(11):2279–85. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31444634

16. Pribenszky C, Nilselid A-MM, Montag M. Time-lapse culture with morphokinetic embryo selection improves pregnancy and live birth chances and reduces early pregnancy loss: a meta-analysis. Reproductive BioMedicine Online. 2017/07/10. 2017 Nov [cited 2020 Jan 13]; 35(5):511–20. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1472648317303073

17. Magdi Y, Samy A, Abbas AM, Ibrahim MA, Edris Y, El-Gohary A, et al. Effect of embryo selection based morphokinetics on IVF/ICSI outcomes: evidence from a systematic review and meta-analysis of randomized controlled trials. Archives of gynecology and obstetrics. 2019/10/30. 2019 Dec; 300(6):1479–90. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31667608

18. Barrie A, Homburg R, McDowell G, Brown J, Kingsland C, Troup S. Examining the efficacy of six published time-lapse imaging embryo selection algorithms to predict implantation to demonstrate the need for the development of specific, in-house morphokinetic selection algorithms. Fertility and Sterility. 2017/01/06. 2017 Mar 1 [cited 2020 Jan 13]; 107(3):613–21. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0015028216630145

19. Alikani M, Go KJ, McCaffrey C, McCulloh DH. Comprehensive evaluation of contemporary assisted reproduction technology laboratory operations to determine staffing levels that promote patient safety and quality care. Fertility and Sterility. 2014; 102(5):1350–6. Available from: http://dx.doi.org/10.1016/j.fertnstert.2014.07.1246

20. Sundvall L, Ingerslev HJ, Breth Knudsen U, Kirkegaard K. Inter- and intra-observer variability of time-lapse annotations. Human Reproduction. 2013/09/26. 2013 Dec 1 [cited 2020 Jan 13]; 28(12):3215–21. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24070998

21. Adolfsson E, Andershed AN. Morphology vs morphokinetics: A retrospective comparison of interobserver and intra-observer agreement between embryologists on blastocysts with known implantation outcome. Jornal Brasileiro de Reproducao Assistida. 2018 Sep 1 [cited 2020 Jan 13]; 22(3):228–37. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29912521

22. 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. Human Reproduction. 2017; 32(2):307–14. https://doi.org/10.1093/humrep/dew330 PMID: 28031323

23. Bormann CL, Ph D, Thirumalaraju P, Tech B, Kanakasabapathy K, Tech M. Consistency and objectivity of automated embryo assessments using deep neural networks. Fertility and sterility. 2020; 113 (4):781–7. https://doi.org/10.1016/j.fertnstert.2019.12.004 PMID: 32228880

24. Giusti A, Corani G, Gambardella L, Magli C, Gianaroli L. Blastomere segmentation and 3D morphology measurements of early embryos from hoffman modulation contrast image stacks. In: 2010 7th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2010—Proceedings. 2010. p. 1261–4.

25. Wang Y, Moussavi F, Lorenzen P. Automated embryo stage classification in Time-Lapse Microscopy Video of Early Human Embryo Development. Medical image computing and computer-assisted intervention: MICCAI. International Conference on Medical Image Computing and Computer-Assisted Intervention. 2013; 16(Pt 2):460–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24579173

26. Kragh MF, Rimestad J, Berntsen J, Karstoft H. Automatic grading of human blastocysts from time-lapse imaging. Computers in Biology and Medicine. 2019/10/15. 2019 Dec 1; 115:103494. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31630027

27. Khosravi P, Kazemi E, Zhan Q, Toschi M, Malmsten JE, Hickman C, et al. Robust Automated Assessment of Human Blastocyst Quality using Deep Learning. bioRxiv. 2018; Available from: https://www.biorxiv.org/content/10.1101/394882v1

28. Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. npj Digital Medicine. 2019 Dec 4; 2(1):21. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31304368

29. Ng NH, McAuley J, Lipton ZC, Gingold JA, Desai N. Predicting embryo morphokinetics in videos with late fusion nets & dynamic decoders. In: 6th International Conference on Learning Representations, ICLR 2018—Workshop Track Proceedings. 2018.

30. Rocha JC, da Silva DLB, dos Santos JGC, Whyte LB, Hickman C, Lavery S, et al. Using artificial intelligence to improve the evaluation of human blastocyst morphology. In: IJCCI 2017—Proceedings of the 9th International Joint Conference on Computational Intelligence. SciTePress; 2017. p. 354–9.

31. Kanakasabapathy MK, Thirumalaraju P, Bormann CL, Kandula H, Dimitriadis I, Souter I, et al. Development and evaluation of inexpensive automated deep learning-based imaging systems for embryology.

Lab on a chip. 2019/11/22. 2019 Dec 21; 19(24):4139–45. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31755505

32. Chavez-Badiola A, Flores-sai A, Mendizabal-ruiz G, Garcia-sanchez R, Drakeley AJ, Garcia-sandoval JP. Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. Scientific Reports. 2020; 10(4394):1–6. https://doi.org/10.1038/s41598-020-61357-9 PMID: 32157183

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. Fertility & Reproduction. 2019 Mar; 01(01):51–6.

34. Kan-Tor Y, Zabari N, Erlich I, Szeskin A, Amitai T, Richter D, et al. Automated Evaluation of Human Embryo Blastulation and Implantation Potential using Deep-Learning. Advanced Intelligent Systems. 2020 Oct; 2(10):2000080. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202000080

35. Tran D, Cooke S, Illingworth PJ, Gardner DK. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Human reproduction. 2019 Jun 4; 34 (6):1011–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31111884

36. Apter S, Ebner T, Freour T, Guns Y, Kovacic B, le Clef N, et al. Good practice recommendations for the use of time-lapse technology. Human Reproduction Open. 2020;1–26.

37. Carreira J, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Vols. 2017-Janua, Proceedings—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. 2017. p. 4724–33. http://arxiv.org/abs/1705.07750

38. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997; 9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735 PMID: 9377276

39. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings. 2015. p. 1–15.

40. Smith LN, Topin N. Super-convergence: very fast training of neural networks using large learning rates. In 2019. p. 36.

41. Goodfellow IJ, Warde-Farley D, Mirza M, Courville A, Bengio Y. Maxout networks. In: 30th International Conference on Machine Learning, ICML 2013. 2013. p. 2356–64.

42. Steyerberg EW. Clinical Prediction Models. Statistics for Biology and Health. 2nd edition. 2019.

43. DeVries T, Taylor GW. Improved Regularization of Convolutional Neural Networks with Cutout. 2017 Aug 15;

44. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020; 42(2):318–27. https://doi.org/10.1109/TPAMI.2018.2858826 PMID: 30040631

45. Kajhøj TQ. The language of embryology is evolving—a guide to understanding time-lapse nomenclature. 2016. https://blog.vitrolife.com/togetheralltheway/a-guide-to-understanding-time-lapse-nomenclature

46. Zhan Q, Ye Z, Clarke R, Rosenwaks Z, Zaninovic N. Direct unequal cleavages: Embryo developmental competence, genetic constitution and clinical outcome. PLoS ONE. 2016 Dec 1; 11(12):1–19. Available from: https://pubmed.ncbi.nlm.nih.gov/27907016/?from_term=Zhan+Q&from_cauthor_id=27907016&from_pos=1

47. Vitrolife. KIDScore D5 decision support tool. 2019. p. 1–2. https://www.vitrolife.com/globalassets/support-documents/tech-notes/technote_kidscore-d5.3_v3_v2.pdf

48. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12(1):77. Available from: http://www.biomedcentral.com/1471-2105/12/77

49. Curchoe CL, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Chavez-Badiola A. Evaluating predictive models in reproductive medicine. Fertility and Sterility. 2020 Nov; 114(5):921–6. Available from: https://doi.org/10.1016/j.fertnstert.2020.09.159

50. Kragh MF, Karstoft H. Embryo selection with artificial intelligence: how to evaluate and compare methods? Journal of Assisted Reproduction and Genetics. 2021 June 38: 1675–1689. Available from: https://link.springer.com/article/10.1007%2Fs10815-021-02254-6

51. Kirkegaard K, Campbell A, Agerholm I, Bentin-Ley U, Gabrielsen A, Kirk J, et al. Limitations of a time-lapse blastocyst prediction model: A large multicentre outcome analysis. Vol. 29, Reproductive BioMedicine Online. Elsevier Ltd; 2014. p. 156–8.

52. Coticchio G, Lagalla C, Sturmey R, Pennetta F, Borini A. The enigmatic morula: mechanisms of development, cell fate determination, self-correction and implications for ART. Human Reproduction Update. 2019 Jul 1; 25(4):422–38. Available from: https://academic.oup.com/humupd/article/25/4/422/5374477

53. Shapiro BS, Daneshmand ST, Garner FC, Aguirre M, Thomas S. Large blastocyst diameter, early blastulation, and low preovulatory serum progesterone are dominant predictors of clinical pregnancy in fresh autologous cycles. Fertility and Sterility. 2008; 90(2):302–9. https://doi.org/10.1016/j.fertnstert.2007.06.062 PMID: 17905239

54. Gardner DK, Lane M, Stevens J, Schlenker T, Schoolcraft WB. Blastocyst score affects implantation and pregnancy outcome: Towards a single blastocyst transfer. Fertility and Sterility. 2000; 73(6):1155–8. https://doi.org/10.1016/s0015-0282(00)00518-5 PMID: 10856474

55. Ahlström A, Westin C, Reismer E, Wikland M, Hardarson T. Trophectoderm morphology: an important parameter for predicting live birth after single blastocyst transfer. Human reproduction. 2011/10/03. 2011 Dec; 26(12):3289–96. Available from: https://pubmed.ncbi.nlm.nih.gov/21972253

56. Gardner DK, Sakkas D. Assessment of embryo viability: the ability to select a single embryo for transfer —a review. Placenta. 2003 Oct; 24 Suppl B:5–12. Available from: https://pubmed.ncbi.nlm.nih.gov/14559024

57. VerMilyea M, Hall JMM, Diakiw SM, Johnston A, Nguyen T, Perugini D, et al. 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. 2020 Apr 2;1–15. Available from: https://academic.oup.com/humrep/advance-article/doi/10.1093/humrep/deaa013/5815143

58. Chavez-Badiola A, Mendizabal-Ruiz G, Flores-Saiffe Farias A, Garcia-Sanchez R, Drakeley AJ. Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Human Reproduction. 2020 Feb 29; 35(2):482–482. Available from: https://academic.oup.com/humrep/article/35/2/482/5717668

59. Kragh MF, Rimestad J, Lassen JT, Berntsen J, Karstoft H. Predicting embryo viability based on self-supervised alignment of time-lapse videos, in IEEE Transactions on Medical Imaging, https://doi.org/10.1109/TMI.2021.3116986 2021. PMID: 34596537

60. Caron M, Goyal P, Misra I, Bojanowski P, Mairal J, Joulin A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. arXiv. 2020;(NeurIPS):1–23.

61. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. ViViT: A Video Vision Transformer. 2021; http://arxiv.org/abs/2103.15691

62. Bertasius G, Wang H, Torresani L. Is Space-Time Attention All You Need for Video Understanding? 2021; http://arxiv.org/abs/2102.05095