Human Reproduction, Vol.34, No.6, pp. 1011–1018, 2018
Advance Access Publication on May 21, 2019 doi:10.1093/humrep/dez064

human reproduction

ORIGINAL ARTICLE *Embryology*

Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer

D. Tran[1,*], S. Cooke[2], P.J. Illingworth[2], and D.K. Gardner[3]

[1]Medical AI, Harrison AI, Barangaroo, NSW, Australia [2]Embryology, IVF Australia, Greenwich, NSW, Australia [3]Embryology, Melbourne IVF, East Melbourne, Victoria, Australia

*Correspondence address. Harrison AI, Barangaroo, New South Wales, Australia. E-mail: Aengus@harrison.ai

Submitted on February 26, 2019; resubmitted on April 7, 2019; editorial decision on April 15, 2019

STUDY QUESTION: Can a deep learning model predict the probability of pregnancy with fetal heart (FH) from time-lapse videos?

SUMMARY ANSWER: We created a deep learning model named IVY, which was an objective and fully automated system that predicts the probability of FH pregnancy directly from raw time-lapse videos without the need for any manual morphokinetic annotation or blastocyst morphology assessment.

WHAT IS KNOWN ALREADY: The contribution of time-lapse imaging in effective embryo selection is promising. Existing algorithms for the analysis of time-lapse imaging are based on morphology and morphokinetic parameters that require subjective human annotation and thus have intrinsic inter-reader and intra-reader variability. Deep learning offers promise for the automation and standardization of embryo selection.

STUDY DESIGN, SIZE, DURATION: A retrospective analysis of time-lapse videos and clinical outcomes of 10 638 embryos from eight different IVF clinics, across four different countries, between January 2014 and December 2018.

PARTICIPANTS/MATERIALS, SETTING, METHODS: The deep learning model was trained using time-lapse videos with known FH pregnancy outcome to perform a binary classification task of predicting the probability of pregnancy with FH given time-lapse video sequence. The predictive power of the model was measured using the average area under the curve (AUC) of the receiver operating characteristic curve over 5-fold stratified cross-validation.

MAIN RESULTS AND THE ROLE OF CHANCE: The deep learning model was able to predict FH pregnancy from time-lapse videos with an AUC of 0.93 [95% CI 0.92–0.94] in 5-fold stratified cross-validation. A hold-out validation test across eight laboratories showed that the AUC was reproducible, ranging from 0.95 to 0.90 across different laboratories with different culture and laboratory processes.

LIMITATIONS, REASONS FOR CAUTION: This study is a retrospective analysis demonstrating that the deep learning model has a high level of predictability of the likelihood that an embryo will implant. The clinical impacts of these findings are still uncertain. Further studies, including prospective randomized controlled trials, are required to evaluate the clinical significance of this deep learning model. The time-lapse videos collected for training and validation are Day 5 embryos; hence, additional adjustment would need to be made for the model to be used in the context of Day 3 transfer.

WIDER IMPLICATIONS OF THE FINDINGS: The high predictive value for embryo implantation obtained by the deep learning model may improve the effectiveness of previous approaches used for time-lapse imaging in embryo selection. This may improve the prioritization of the most viable embryo for a single embryo transfer. The deep learning model may also prove to be useful in providing the optimal order for subsequent transfers of cryopreserved embryos.

STUDY FUNDING/COMPETING INTEREST(S): D.T. is the co-owner of Harrison AI that has patented this methodology in association with Virtus Health. P.I. is a shareholder in Virtus Health. S.C., P.I. and D.G. are all either employees or contracted with Virtus Health. D.G. has received grant support from Vitrolife, the manufacturer of the Embryoscope time-lapse imaging used in this study. The equipment and time for this study have been jointly provided by Harrison AI and Virtus Health.

Key words: artificial intelligence / deep learning / neural network / embryo selection / time-lapse

Introduction

The advent of more physiological culture conditions for the human embryo, conceived through IVF, has led to the routine culture and transfer of embryos at the blastocyst stage (Gardner *et al.*, 1998; Biggers and Racowsky, 2002). Transfer of a single blastocyst can avoid the many adverse medical conditions for mother and baby associated with multiple gestations (Adashi *et al.*, 2003; Sullivan *et al.*, 2012).

However, our ability to select the optimum embryo for transfer has changed very little since the birth of Louise Brown about 40 years ago (Steptoe and Edwards, 1978). From the beginnings of IVF, it was noted that the rate of embryo development was associated with transfer outcome (Edwards *et al.*, 1984). Subsequently, elegant grading systems have been developed for each successive stage of preimplantation human embryo development (Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology, 2011; Gardner and Balaban, 2016), and each one has been able to relate aspects of discrete stages of development at specific times to viability. In the case of the blastocyst, an alpha numeric system has been in place since the turn of the century, which takes into account both inner cell mass development, as well as the trophectoderm, and its biological activity (measured indirectly through the expansion of the blastocoel cavity itself) (Schoolcraft *et al.*, 1999; Gardner *et al.*, 2000). Although such an approach has proved particularly effective in selecting embryos for transfer, data on the metabolic activity of the human blastocyst indicate that Day 5 morphology alone is not the sole predictor of embryonic viability (Gardner *et al.*, 2015), with glucose consumption being positively correlated with the ability of a blastocyst to give rise to a pregnancy (Gardner *et al.*, 2011)

Prior to the introduction of time-lapse technologies, all embryo assessments were restricted to specific time points during the first 5 days of life. Clearly, the majority of developmental events were not being captured, and the use of time-lapse to image the developing embryo every few minutes has confirmed that several key features of human embryo development were being missed, such as direct cleavage (Rubio *et al.*, 2012), or were simply not being quantitated, such as the time taken to initiate and complete cavitation. This latter point is of great physiological interest, as it appears tied to both metabolic activity and embryonic ploidy (Desai *et al.*, 2018). In an attempt to glean more predictive potential from time-lapse images, several algorithms have been created (Motato *et al.*, 2016; Petersen *et al.*, 2016; Fishel *et al.*, 2018), which take into account the annotated time at which key morphological events occur. Data to date have suggested that this approach may have value in embryo selection, and its efficacy has been studied in randomized trials. Results have been variable (Rubio *et al.*, 2014; Goodman *et al.*, 2016), but accumulating data infer that key kinetic events are associated with important biological information (Desai *et al.*, 2018).

However, a current limitation of the analysis of time-lapse images through these algorithms was that it was unfeasible to capture the full temporal and spatial richness of time-lapse video within a few morphokinetic parameters. As such, time-lapse data are currently under-utilized in making predictions about clinical outcomes. Such morphokinetic algorithms also rely on the embryologist to manually annotate the morphological features and morphokinetic timing data. These parameters can then be used as input for statistical or machine-learned scoring tools. Unfortunately, time-lapse annotation and grading is a subjective process with intrinsic inter-reader and intra-reader variability (Venetis *et al.*, 2017) that will impact the performance of the downstream scoring tools. What determines fetal heart (FH) pregnancy was likely not a simple correlation between a few known features, but rather a complex interaction between many factors across both the temporal and spatial dimension (information acquired through time-lapse microscopy), which may or may not have been identified yet.

Deep learning is a subfield of machine learning that is based on learning hierarchical knowledge from data rather than rule-based programming. Deep learning models are inspired by the biological nervous system in the way that information is processed through interconnecting neurons that are many layers deep (Kim, 2016). Recent advancement in learning algorithms, together with the explosion of electronic medical data and the increase in computational processing power, has seen an explosion of application of artificial intelligence in several aspects of human ART (Bui *et al.*, 2017; Blank *et al.*, 2019; Curchoe and Bormann, 2019). A deep learning algorithm can directly analyse the entire raw time-lapse video without the need for annotated parameters, making use of every data point collected from time-lapse to predict the probability of FH pregnancy. The objective of the study was to investigate the hypothesis that a deep learning model named IVY is a valid tool for the prediction of the implantation potential of human preimplantation embryos.

Materials and Methods

Data collection

This study was carried out across eight IVF laboratories in four countries, IVFAustralia (Sydney, Australia), IVFAustralia (Canberra, Australia), Hunter IVF (Newcastle, Australia), Melbourne IVF (Melbourne, Australia), Queensland Fertility Group (Brisbane, Australia) SIMS IVF (Dublin, Ireland), Complete Fertility Centre (Southampton, UK) and Aagard Fertility (Aarhus, Denmark). Each clinic used its own approach to superovulation, egg collection and embryo transfer as shown in Table I. All embryos were cultured in the Embryoscope^{TM} or Embryoscope Plus^{TM} (Vitrolife, Copenhagen, Denmark). The study was a retrospective analysis of the videos obtained from fresh embryos that were fertilized and cultured in a time-lapse incubator in these laboratories from January 2014 to December 2018.

All embryos cultured to the blastocyst stage in a time-lapse incubator in these laboratories during the time period were studied, regardless of their stage, grade, ploidy, fertilization status and culture method. No embryos cultured in a time-lapse incubator were excluded from the analysis. As a result, the study group included embryos that resulted from fresh oocyte retrieval, vitrified oocyte warming and oocyte donation. Embryos that underwent embryo biopsy for preimplantation genetic testing were included. Embryos that were cryopreserved for later use were studied and the later outcome from the thawed embryo transfer was included in the analysis. Patients from all demographic groups and medical history were included. The mean age was 35.6 years (age range, 22–50 years). No patients were excluded.

The study included 1835 unique treatment cycles from 1648 individual patients. On average, there were 7.9 embryos per treatment. Twenty-eight percent of the embryos transferred were part of a

Table I Number of embryos, patient ages and culture media used in each laboratory.

Laboratory number	Laboratory	Number of embryos studied	Mean age	Age range	Media used in the laboratory
1	IVFAustralia (Sydney, Australia)	1264	36.8	23–46	Vitrolife, Sequential; Vitrolife, Single Stage (G-TL™)
2	IVFAustralia (Canberra, Australia)	150	34.2	24–44	Sage, Sequential
3	Hunter IVF (Newcastle, Australia)	632	34.8	25–43	Vitrolife, Single Stage (G-TL™)
4	Melbourne IVF (Melbourne, Australia)	758	36.6	30–45	Vitrolife, Single Stage (G-TL™)
5	Queensland Fertility Group (Brisbane, Australia)	3827	35.6	22–50	Sage, Sequential; COOK, Sequential; Vitrolife, Single Stage (G-TL™)
6	SIMS IVF (Dublin, Ireland)	1454	35.9	25–46	Vitrolife, Single Stage (G-TL™)
7	Complete Fertility Centre (Southampton, UK)	915	34.7	24–44	Vitrolife, Sequential; Vitrolife, Single Stage (G-TL™)
8	Aagard Fertility (Aarhus, Denmark)	1683	34.2	24–44	Sage 1-step

Table II Classification of the outcome of each embryo for training of the deep learning system.

Classification	Outcome
POSITIVE for each embryo involved	FH observed on ultrasound after 7 weeks gestation following a single embryo transfer or multiple FH observed equal to the number of embryos transferred
NEGATIVE for each embryo involved	No pregnancy occurred or no FH was observed on ultrasound after 7 weeks gestation following transfer or embryo discarded because of a failed or abnormal fertilization, grossly abnormal morphology or aneuploidy from preimplantation genetic testing
UNKNOWN for each embryo involved	Multiple embryos transferred and FH(s) seen but the number is fewer than the number transferred
PENDING	Embryo in storage and not yet used

FH, Fetal heart.

multiple transfer cycle. Twenty-nine percent of the embryos transferred were frozen embryos, and the remainder were fresh embryo transfers. The resulting outcomes for the embryos are shown in Fig. 1.

Deidentified videos, as well as patient age and the outcome for each embryo, were provided to the researchers by each participating laboratory. All other patient information was held confidentially by the participating laboratory, and no identifying information was made available to the researchers. The outcomes of the analysis were not reported back to the participating laboratories.

The clinical outcome for each embryo was classified as per Table II. The principal endpoint was an FH pregnancy that was defined by the observed presence of an FH on ultrasound at or beyond 7 weeks gestation. The entire dataset includes the 8836 embryos coded as positive or negative, which were used for training and testing of the deep learning model.

Training of the deep learning model

The training process began by randomly partitioning the entire dataset into an 80% training dataset and a 20% testing dataset. The deep learning model named IVY was only trained on the embryos in the training set and subsequently used to make predictions on the held-out testing dataset to estimate its predictive power.

IVY is a feed forward deep learning model that takes nothing but the raw time-lapse video sequence as input and produces a confidence score ranging from 0 to 1 as output that represents the model's confidence that the input embryo video will lead to an FH pregnancy. The behaviour characteristic of this deep learning model was dictated by a large set of weights that were randomly initialized at the beginning of training; hence, the model starts out making random predictions.

During the training process, batches of time-lapse videos were randomly sampled from the training dataset. The deep learning model then attempted to make predictions on these time-lapse videos and produce confidence scores for each embryo. These predictions were compared with the known target outcomes (0 for negative and 1 for positive) to compute the difference known as the 'loss', which represents how far off the predictions was compared to the known outcomes.

The loss value was then used to compute the incremental updates on all of the model's weights to improve its prediction, a process known

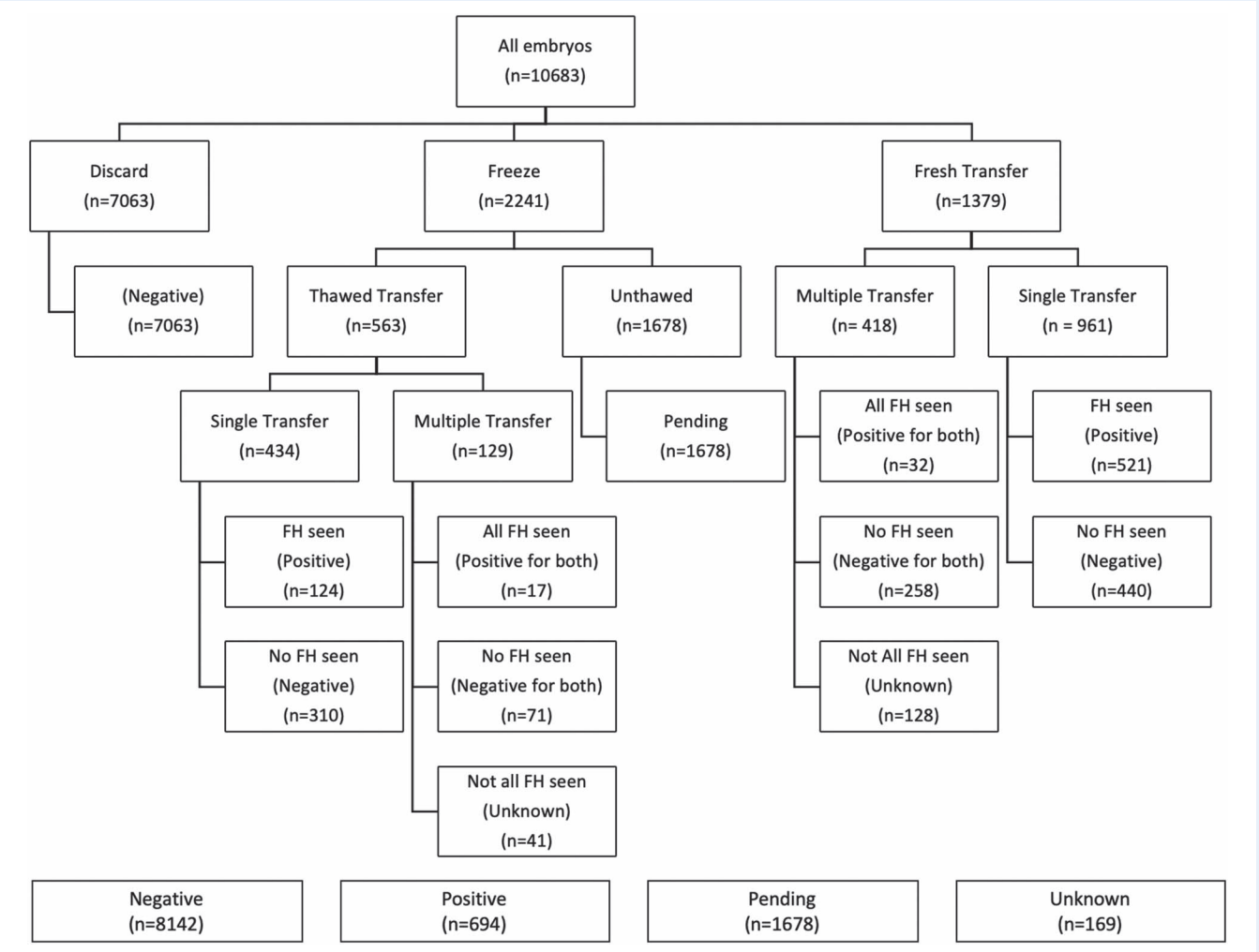

Figure 1 The outcomes of the embryos being studied. FH, Fetal heart.

as backpropagation. This process was repeated for many thousands of updates to drive down the loss value. As the training progressed, the loss value was continuously reduced closer to 0, which translate to predictions that were closer to the actual FH pregnancy outcome. The training set was looped over randomly for 20 epochs. Each epoch represents a single pass over the dataset. Once the loss value had plateaued, and no more improvement could be made, the training was stopped. The trained model and all its weights were then saved to make predictions against the unseen embryos in the testing dataset. The predictions were compared with the known FH pregnancy outcome from the testing dataset to obtain the performance characteristic of the trained model.

ROC

The performance characteristic of IVY was calculated using the receiver operating characteristic (ROC) curve generated by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across all possible thresholding values using the predicted confidence score compared to the actual FH pregnancy outcome. Sensitivity and specificity values can be calculated by selecting a specific thresholding value. A lower threshold value will yield a higher sensitivity and lower specificity while a higher threshold value will yield a lower sensitivity and higher specificity. The nature of this trade-off can be summarized by calculating the area under the curve (AUC) of the ROC curve.

AUC ranged from 0.5 to 1.0, which represents the predictive power of a binary classifier. An AUC of 0.5 represents completely random choices while an AUC of 1 represents perfect discrimination. The higher the AUC, the more favourable the trade-off between sensitivity and specificity. The quantitative value of AUC can also simply be interpreted as the probability that the binary classifier will score a randomly selected positive embryo higher than a randomly selected negative embryo. As a result, AUC was the most appropriate benchmark for a binary classifier's ability to rank embryos according to their likelihood of creating an FH pregnancy.

5-fold cross-validation

To increase the robustness of the performance estimation we performed 5-fold stratified cross-validation (Kuhn and Johnson, 2013). During this process, the entire dataset was randomly partitioned

into five subsets of equal size such that each subset maintained the same prevalence of positive embryos. We subsequently trained five separate deep learning models from scratch on four out of the five subsets and performed validation on the fifth hold-out subset. For each of the training-testing runs, we calculated the AUC on the testing dataset as described above. The final AUC was reported as the mean AUC over five separate training-testing runs. Cross-validation enables the estimate a more reliable performance metric. This methodology lowers the risk of overestimating or underestimating the model true performance characteristic by sampling a uniquely favourable or challenging testing dataset by chance.

Eight laboratories hold-out validation

To investigate the transferability of the deep learning model across different laboratory environments and patient demographics, we also performed an eight-laboratory hold-out validation test (Kuhn and Johnson, 2013). In this approach, the entire dataset was partitioned into eight cohorts according to their laboratory of origin. We subsequently trained another eight deep learning models from scratch by holding out embryos from each of the eight laboratories for validation and performed training on the other seven laboratories' videos. We then used each of the eight trained models to calculate the AUC on the embryos from the corresponding hold-out laboratories to estimate its predictive power on the untrained lab.

Ethical approval

Under the Australian National Statement on Ethical Conduct in Human Research (National Health and Medical Research Council, 2015), this project was classified as negligible risk and was exempted from ethical review.

Results

ROC

Analysis of the ROC is shown in Fig. 2. The resulting AUC of IVY to predict FH pregnancy on the testing dataset was 0.93 (95% CI 0.92–0.94).

5-fold stratified cross-validation

The results of the 5-fold cross-validation are shown in Table III. The mean AUC over 5-fold stratified cross-validation was 0.93. The AUC was reproducible across five separate training-validation runs.

Eight laboratories hold-out validation

The results of the eight laboratories hold-out validation are shown in Supplementary Table S1. The validation AUC from each laboratory was very similar indicating excellent transferability across the different laboratory process and clinical environments.

Discussion

The data presented demonstrate that deep learning can predict, with a high degree of probability and reproducibility, the likelihood of a transferred embryo implanting and developing as far as FH. The AUC

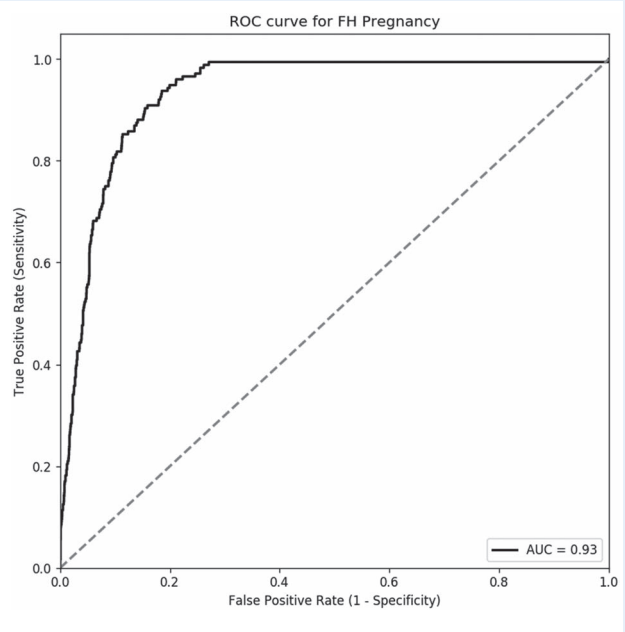

Figure 2 ROC curve for prediction of FH pregnancy on the testing dataset by IVY. ROC, Receiver operating characteristic; AUC, area under the curve.

finding presented here categorizes the diagnostic test quality of IVY as excellent (Šimundić, 2009). Our findings suggest that the combination of time-lapse imaging and deep learning provides an assessment of the embryo viability that is likely to be significantly better than previous algorithm-based approaches.

Previous approaches to morphokinetic analysis (Meseguer *et al.*, 2012; Conaghan *et al.*, 2013; Basile *et al.*, 2014; Milewski *et al.*, 2016; Fishel *et al.*, 2018; Liu *et al.*, 2018) have been based on algorithms that have studied known events in embryological development, such as rapid early cleavage and blastulation, and applied scores to these known events. Deep learning takes an entirely different approach, where time-lapse videos are objectively reviewed with no assumptions at all being made about the significance, or otherwise, of different events in early embryo development.

A number of studies have examined the predictive power of morphokinetic algorithms. A comparison of six previously published embryo selection algorithms to a single set of blastocysts (Barrie *et al.*, 2017) found that none of the algorithms had an AUC of greater than 0.65. A study from our own group (Storr *et al.*, 2015) carried out a logistic regression of four different approaches. They found the AUC on ROCs curve varied from 0.585 to 0.748. A further study (Liu *et al.*, 2018) observed a wide range of predictive powers from 0.509 to 0.762, depending on the algorithm tested.

Previous studies of the predictive capacity of morphokinetics on Day 2 embryos (Ahlstrom *et al.*, 2016; Milewski *et al.*, 2016) give comparable results and have suggested that an embryologist approach gives an AUC of 0.74 (Ahlstrom *et al.*, 2016) while morphokinetics gives an AUC of 0.67 (Ahlstrom *et al.*, 2016) and 0.70 (Milewski *et al.*, 2016).

The inbuilt software in the Embryoscope system is the KIDscore (Petersen *et al.*, 2016) whereby embryos are allocated to one of five

Table III Results of the 5-fold cross-validation analysis.

	Fold 1 (*n* = 1767)	Fold 2 (*n* = 1767)	Fold 3 (*n* = 1767)	Fold 4 (*n* = 1767)	Fold 5 (*n* = 1768)	AUC
1	Test	Train	Train	Train	Train	0.93
2	Train	Test	Train	Train	Train	0.93
3	Train	Train	Test	Train	Train	0.92
4	Train	Train	Train	Test	Train	0.94
5	Train	Train	Train	Train	Test	0.93
					Mean AUC	**0.93**

Mean AUC, The mean area under the curve across 5 cross-validation steps.

categories (KID1–KID5) based on early cleavage events. This approach has previously been studied and found to surpass the predictive value of 0.745 for morphology based on accepted criteria (Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology, 2011).

Two recent reports have added understanding to this field. Blank *et al.* (2019) has studied the use of machine learning to combine known morphokinetic algorithms, similar to the above, with clinical predictors of IVF outcome such as parental age, Anti-Mullerian hormone concentration and past pregnancies. Although it is also an application of machine learning to the problem of embryo selection, the overall approach is quite different from the deep learning methodology of the present study with a correspondingly lower AUC (0.74) on ROC analysis.

Khosravi *et al.* (2019) have used deep learning to analyse blastocyst images derived from time-lapse imaging and were able to demonstrate a very high ROC (0.98) in being able to predict the embryologist assessment of the embryo. However, unlike our model, this group was unable to demonstrate any direct predictive value for pregnancy. This model differs from our model in being based on only a limited number of images. Our deep learning model studies the whole video and it is likely to be this that gives it the capacity to identify the likelihood of pregnancy with a high degree of predictability.

Clearly, the clinically significant endpoint for any fertility intervention is live birth per cycle started. In this study, we have used FH as the assessment of clinical pregnancy per embryo transferred, as the approach being evaluated is a prediction of the implantation rate for each individual embryo. We acknowledge that the FH rate is clearly only a proxy for live birth, and future clinical studies will need to evaluate the impact on live birth rate. However, at this stage of assessing a rapidly developing technology, the use of live birth as an endpoint would be impractical.

One further limitation of this study that should be noted is that the deep learning system was trained on data from two specific incubator systems (Embryoscope^{TM} and Embryoscope Plus^{TM}, Vitrolife) and the applicability to other time-lapse incubator systems remains unclear. However, the different laboratories involved in this study have heterogeneous patient characteristics, apply a wide range of clinical IVF stimulation regimes and use a variety of different culture media. The use of the eight laboratories hold-out validation demonstrates that the predictive value of the deep learning approach is robust despite a wide range of clinical settings of IVF in different parts of the world.

Previous randomized controlled trials of morphokinetics applied to time-lapse images have not provided convincing evidence of any clinical benefit from such approaches. Some studies have not found any benefit of time-lapse images in improving success rates (Ahlstrom *et al.*, 2016; Goodman *et al.*, 2016), while others have suggested a benefit in success rate (Rubio *et al.*, 2014). Further, it is difficult to separate the different effects of the morphokinetic analysis from varying laboratory conditions. Given the relatively low predictive value with the algorithm-based approaches, a paucity of clinical efficacy may not be entirely surprising. The significantly improved predictive value of the deep learning approach such as IVY suggests that further clinical prospective studies based on a deep learning approach are needed to investigate the possibility of a clinical benefit from this approach.

As in all of IVF, the cumulative likelihood of a successful pregnancy for one cycle started is a consequence of a number of factors including the clinical context, the stimulation approach, the culture system used and the clinic's own inherent success rates. Given modern cryopreservation capacity, the cumulative success rate would be very unlikely to be affected by the ability to predict the implantation rate for a particular embryo. However, the capacity to predict the likelihood of an embryo transfer successfully will result in the embryo with the highest developmental potential being selected first. This would not necessarily affect the cumulative likelihood of obtaining a pregnancy from a batch of embryos cultured from an IVF cycle, but will shorten the time to pregnancy.

The development of deep learning systems may also offer other benefits in the IVF laboratory. Deep learning may increase laboratory efficiency, particularly compared to the time-consuming work involved in annotating morphokinetic parameters as well as eliminating inter-observer variation between embryologists. These potential benefits merit investigation through cost analysis studies.

The data presented here open up the exciting area of investigation of the relationship between the critical features of time-lapse imaging that the deep learning identifies as predictive of implantation and the cellular and physiological events of preimplantation embryo development. Considering that the model operates at raw pixel level across time, macroscopic features such as embryo/cell shape, size and finer features such as texture and movement pattern can all be learned through the training data. The model was trained *tabula rasa* and decides for itself the significance or otherwise of all these features (or the complex interaction between these features) and assigns predictive weighting

accordingly. However, the exact logic utilized by the model, in making its decisions, remains an active area of investigation and will be a topic for future study.

Further, the model presented utilizes the entire video to inform its prediction of the study endpoint in one single pass, which is distinctly different to how embryologist analyse time-lapse video from frame to frame. There may be time intervals in the 5 days of development that are more important than others but this will require further exploration. Future studies to correlate IVY's prediction with other parameters of embryo physiology, such as metabolic activity and chromosomal constitution, will be of value in advancing our understanding of preimplantation embryo development.

One major question is how these predictive calculations relate to other clinical factors that are known to predict embryo outcomes such as female age and previous reproductive history. The findings presented in this study, quite deliberately, result from the study of each individual embryo from a heterogeneous group of patients and the predictive value that has been derived is independent of age, or any other clinical factors. Further studies investigating the prognostic relationship between known clinical factors and deep learning prediction will therefore be of great value.

Conclusion

In summary, these data demonstrate the potential for deep learning to contribute to clinical IVF in the same way as it has contributed to other areas of human health (Patel *et al.*, 2009). This study is a retrospective analysis that has demonstrated an effective means of IVY in predicting implantation rate. Further detailed prospective clinical studies are now underway to investigate the clinical impact and cost-effectiveness of this development.

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Acknowledgements

We acknowledge the contributions of the laboratory staff and directors for feedback, project compliance and video extraction at IVFAustralia-Eastern Suburbs, Hunter IVF, IVFAustralia-Canberra, Queensland Fertility Group, Melbourne IVF, Aagaard Fertility, SIMS Ireland and Complete Fertility Centre.

Authors' roles

D.T. was responsible for carrying out the deep learning analysis. S.C. was responsible for coordinating the study and assembling the time-lapse videos. P.I. and D.G. contributed to the analysis of the project and the manuscript drafting.

Funding

Vitrolife (to D.G.); Virtus Health (to D.T, P.I and S.C).

Conflict of interest

D.T. is a co-owner of Harrison AI that has patented this methodology in association with Virtus Health. P.I. is a shareholder in Virtus Health. S.C., P.I. and D.G. are all either employees or contracted with Virtus Health. The equipment and time for this study have been jointly provided by Harrison AI and Virtus Health.

References

Adashi EY, Barri PN, Berkowitz R, Braude P, Bryan E, Carr J, Cohen J, Collins J, Devroey P, Frydman R *et al.* Infertility therapy-associated multiple pregnancies (births): an ongoing epidemic. *Reprod Biomed Online* 2003;**7**:515–542.

Ahlstrom A, Park H, Bergh C, Selleskog U, Lundin K. Conventional morphology performs better than morphokinetics for prediction of live birth after day 2 transfer. *Reprod Biomed Online* 2016;**33**:61–70.

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

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. *Fertil Steril* 2017;**107**:613–621.

Basile N, Vime P, Florensa M, Aparicio Ruiz B, Garcia Velasco JA, Remohi J, Meseguer M. The use of morphokinetics as a predictor of implantation: a multicentric study to define and validate an algorithm for embryo selection. *Hum Reprod* 2014;**30**:276–283.

Biggers JD, Racowsky C. The development of fertilized human ova to the blastocyst stage in KSOMAA medium: is a two-step protocol necessary? *Reprod Biomed Online* 2002;**5**:133–140.

Blank C, Wildeboer RR, DeCroo I, Tilleman K, Weyers B, de SP, Mischi M, Schoot BC. Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. *Fertil Steril* 2019;**111**:318–326.

Bui TTH, Belli M, Fassina L, Vigone G, Merico V, Garagna S, Zuccotti M. Cytoplasmic movement profiles of mouse surrounding nucleolus and not-surrounding nucleolus antral oocytes during meiotic resumption. *Mol Reprod Dev* 2017;**84**:356–362.

Conaghan J, Chen AA, Willman SP, Ivani K, Chenette PE, Boostanfar R, Baker VL, Adamson GD, Abusief ME, Gvakharia M. Improving embryo selection using a computer-automated time-lapse image analysis test plus day 3 morphology: results from a prospective multicenter trial. *Fertil Steril* 2013;**100**:412–419.

Curchoe CL, Bormann CL. Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018. *J Assist Reprod Genet* 2019;1–10.

Desai N, Goldberg JM, Austin C, Falcone T. Are cleavage anomalies, multinucleation, or specific cell cycle kinetics observed with time-lapse imaging predictive of embryo developmental capacity or ploidy? *Fertil Steril* 2018;**109**:665–674.

Edwards RG, Fishel SB, Cohen J, Fehilly CB, Purdy JM, Slater JM, Steptoe PC, Webster JM. Factors influencing the success of in vitro fertilization for alleviating human infertility. *J In Vitro Fert Embryo Transf* 1984;**1**:3–23.

Fishel S, Campbell A, Montgomery S, Smith R, Nice L, Duffy S, Jenner L, Berrisford K, Kellam L, Smith R *et al.* Time-lapse imaging algorithms rank human preimplantation embryos according to the probability of live birth. *Reprod Biomed Online* 2018;**37**:304–313.

Gardner DK, Balaban B. Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'OMICS': is looking good still important? *Basic Sci Reprod Med* 2016;**22**:704–718.

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

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

Gardner DK, Meseguer M, Rubio C, Treff NR. Diagnosis of human preimplantation embryo viability. *Hum Reprod Update* 2015;**21**:727–747.

Gardner DK, Wale PL, Collins R, Lane M. Glucose consumption of single post-compaction human embryos is predictive of embryo sex and live birth outcome. *Hum Reprod* 2011;**26**:1981–1986.

Goodman LR, Goldberg J, Falcone T, Austin C, Desai N. Does the addition of time-lapse morphokinetics in the selection of embryos for transfer improve pregnancy rates? A randomized controlled trial. *Fertil Steril* 2016;**105**:275–285.

Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, Sigaras A, Lavery S, Cooper LAD, Hickman C *et al.* Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. *NPJ Digit Med* 2019;**2**:21.

Kim KG. Book review: Deep Learning. *Healthc Inform Res* 2016;**22**:351.

Kuhn M, Johnson K. *Applied Predictive Modeling*. New York: Springer, 2013

Liu Y, Feenan K, Chapple V, Matson P. Assessing efficacy of day 3 embryo time-lapse algorithms retrospectively: impacts of dataset type and confounding factors. *Hum Fertil* 2018;1–9.

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**:1481–1489.

Milewski R, Milewska AJ, Kuczyńska A, Stankiewicz B, Kuczyński W. Do morphokinetic data sets inform pregnancy potential? *J Assist Reprod Genet* 2016;**33**:357–365.

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

National Health and Medical Research Council. *National Statement on Ethical Conduct in Human Research 2007 (Updated May 2015)*. 2015; https://nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research (14th April 2019, date last accessed).

Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, Abu-Hanna A. The coming of age of artificial intelligence in medicine. *Artif Intell Med* 2009;**46**:5–17.

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

Rubio I, Galán A, Larreategui Z, Ayerdi F, Bellver J, Herrero J, Meseguer M. Clinical validation of embryo culture and selection by morphokinetic analysis: a randomized, controlled trial of the EmbryoScope. *Fertil Steril* 2014;**102**:1287–1294.

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

Schoolcraft WB, Gardner DK, Lane M, Schlenker T, Hamilton F, Meldrum DR. Blastocyst culture and transfer: analysis of results and parameters affecting outcome in two in vitro fertilization programs. *Fertil Steril* 1999;**72**:604–609.

Šimundić A-M. Measures of diagnostic accuracy: basic definitions. *EJIFCC* 2009;**19**:203–211.

Steptoe PC, Edwards RG. Successful birth after IVF. *Lancet* 1978;**312**:0.

Storr A, Venetis CA, Cooke S, Susetio D, Kilani S, Ledger W. Morphokinetic parameters using time-lapse technology and day 5 embryo quality: a prospective cohort study. *J Assist Reprod Genet* 2015;**32**:1151–1160.

Sullivan EA, Wang YA, Hayward I, Chambers GM, Illingworth P, McBain J, Norman RJ. Single embryo transfer reduces the risk of perinatal mortality, a population study. *Hum Reprod* 2012;**27**:3609–3615.

Venetis CA, Cooke S, Kilani S, Ledger W, Storr A. Inter-observer and intra-observer agreement between embryologists during selection of a single Day 5 embryo for transfer: a multicenter study. *Hum Reprod* 2017;**32**:307–314.