Journal of Assisted Reproduction and Genetics (2019) 36:591–600
https://doi.org/10.1007/s10815-019-01408-x

ASSISTED REPRODUCTION TECHNOLOGIES

Artificial intelligence and machine learning for human reproduction and embryology presented at ASRM and ESHRE 2018

Carol Lynn Curchoe[1] · Charles L. Bormann[2]

Received: 15 October 2018 / Accepted: 15 January 2019 / Published online: 28 January 2019

Abstract

Sixteen artificial intelligence (AI) and machine learning (ML) approaches were reported at the 2018 annual congresses of the American Society for Reproductive Biology (9) and European Society for Human Reproduction and Embryology (7). Nearly every aspect of patient care was investigated, including sperm morphology, sperm identification, identification of empty or oocyte containing follicles, predicting embryo cell stages, predicting blastocyst formation from oocytes, assessing human blastocyst quality, predicting live birth from blastocysts, improving embryo selection, and for developing optimal IVF stimulation protocols. This represents a substantial increase in reports over 2017, where just one abstract each was reported at ASRM (AI) and ESHRE (ML). Our analysis reveals wide variability in how AI and ML methods are described (from not at all or very generic to fully describing the architectural framework) and large variability on accepted dataset sizes (from just 3 patients with 16 follicles in the smallest dataset to 661,060 images of 11,898 human embryos in one of the largest). AI and ML are clearly burgeoning methodologies in human reproduction and embryology and would benefit from early application of reporting standards.

Keywords Artificial intelligence · Machine learning · Human reproduction · Embryology · ASRM · ASHRE

Introduction

Artificial intelligence (AI) and machine learning (ML) are quickly gaining traction in human reproduction and embryology. In just 1 year, published abstracts at the annual proceedings of the American Society of Reproduction [1, 2] and European Society of Human Reproduction and Embryology [3, 4] increased seven-fold (Table 1).

Despite advances in personalized ovarian stimulation, extended embryo culture, pre-implantation genetic testing, and embryo selection, on average, only one-third of all IVF cycles result in a pregnancy [5]. This represents a significant problem that AI and ML can be leveraged against, as we bear down on the holy grail of our industry: *short time to pregnancy through improved IVF cycle efficiency (reduction of failed retrievals or transfers and miscarriages), from replacement of a single, euploid embryo resulting in a healthy, live-birth.*

Machine learning is based on the idea that we can build machines to process data and learn on their own, without our constant supervision. Machine learning is a way of achieving artificial intelligence. ML algorithms use statistics to find patterns in massive amounts of data. They then use those patterns to make predictions for example, Internet search engines, social media news feeds, digital assistants like Siri and Alexa, Spotify, Netflix, Amazon, and YouTube, GPS navigation, and much more are all powered by AI technology. A simple way to think of AI is in terms of the senses: can it recognize what it sees, respond (sensibly) to what it hears or reads, move in response to what it sees or hears, or "reason", i.e., make decisions based on data? Then, most likely, it is powered by machine learning AI systems.

Over 20 years ago, Kaufmann et al. proposed IVF for infertility care using Cortex Pro neural network software consisting of just four inputs (yes/no freezing, age, number of eggs recovered, number of embryos transferred), one hidden layer of four nodes and one output. Total predictive power as was limited to 59% [6]. Progress toward this advancement of an IVF AI has been slow. Complex (and diverse) datasets,

✉ Carol Lynn Curchoe
carolcurchoe@32atps.com

[1] San Diego Fertility Center, 11425 El Camino Real, San Diego, CA 92130, USA

[2] Department of Obstetrics and Gynecology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Table 1 Artificial intelligence and machine learning abstracts 2017 and 2018

#	Authors	Title	Number of samples	Algorithms and/or architecture	Results
American Society for Reproductive Medicine 2018					
1	J. Kort, D. Meyer, N.Y. Chen, V.L. Baker, J.Y. Huang, D. Camarillo, B. Behr	Predicting blastocyst formation from oocyte mechanical properties: a comparison of a machine learning classifier with embryologist morphological assessment	Thirty-four patients and their corresponding 773 oocytes—of which 213 oocytes underwent measurement and 560 did not.	Zener bulk mechanical model to obtain mechanical parameters: k0, k1, tau, eta0, and eta1.	The machine learning classifier using age, k1, and eta1 was able to correctly predict blastocyst development with a positive predictive value of 80% (95% CI 60.45% to 91.28%) and a negative predictive value of 63.8% (95% CI 53.42% to 73.18%).
2	N. Zaninovic, P. Khosravi, I. Hajirasouliha, J.E. Malmsten, E. Kazemi, Q. Zhan, M. Toschi, O. Elemento, Z. Rosenwaks	Assessing human blastocyst quality using artificial intelligence (AI) convolutional neural network (CNN)	50,392 images from 10,148 embryos cultured in TLM system.	Deep neural networks using Google's Inception (V1) architecture	97.52% accuracy discriminating between poor and good blastocyst groups. 90.6% accuracy for true positive (implanted and good quality BL) and 89.6% accuracy for true negative (non-implanted and poor quality BL). The accuracy of good quality with a negative outcome was 14.63%, while that of poor quality with a positive outcome was 44.78%.
3	N. Zaninovic, C.J. Rocha, Q. Zhan, M. Toschi, J. Malmsten, M. Nogueira, M. Meseguer, Z. Rosenwaks, C. Hickman	Application of artificial intelligence technology to increase the efficacy of embryo selection and prediction of live birth using human blastocysts cultured in a time-lapse incubator	303 ICSI embryos associated with live births resulting from single blastocyst transfers, and 386 TLM images of embryos at 111.5 h post ICSI.	An artificial neural network (ANN) architecture associated with a genetic algorithm.	83% overall accuracy of predicting live birth by AI based on morphokinetic data. The overall accuracy of live birth by AI using image analysis was 85%.
4	K. Iwata, M. Sato, I. Matsumoto, T. Shimura, K. Yumoto, A. Negami, Y. Mio	Deep learning based on images of human embryos obtained from high-resolution time-lapse cinematography for predicting good-quality embryos	118 Normally fertilized embryos from ICSI and Conventional IVF	Deep learning was performed with the Keras neural network library.	94% correct answers for the training dataset and 70% for the validation dataset. Good-quality embryos at the four-cell stage were predictively determined using 31 images captured hourly for 30 h. After 50 learning sessions, correct answers were achieved 92% for the training dataset and 80% for the validation dataset.
5	A. Tran, S. Cooke, P.J. Illingworth, D.K. Gardner	Artificial intelligence as a novel approach for embryo selection	2000 embryos were transferred, 73% of which were single ET, with 670 embryos resulting in a fetal heartbeat (FH). 8389 embryos had known FH outcome data and were used for training or validation.	The AI is a deep neural network trained with a large multicentre, multinational dataset of TL videos to perform binary classification task of predicting FH outcome.	Mean Area Under the Curve (AUC) from 5-fold cross-validation was 0.93, 95% CI [0.92, 0.94] for predicting FH outcome on the validation set; indicating that 93% of the time, the AI will score embryos that lead to a FH higher than embryos that will not.
6	P. Thirumalaraju, C.L. Bormann, M. Kanakasabapathy, F.	Automated sperm morphology testing using artificial intelligence	3500 images of sperm acquired using clinical benchtop microscopes.	Transfer learning with a deep-convolutional network after replacing the final classification layer	371 of 415 individual sperm images were correctly identified (~89%) based on annotations obtained from technicians. The network's

Table 1 (continued)

#	Authors	Title	Number of samples	Algorithms and/or architecture	Results
	Doshi, I. Souter, I. Dimitriadis, H. Shafiee			and retraining it with a dataset of sperm images.	performance in assessing semen samples for sperm morphology was tested. The network performed with an accuracy of 100% in identifying all abnormal and normal samples ($n = 9$) in comparison to the national average reported by the AAB. The sensitivity and specificity of the network was 100% with positive and negative predictive values of 100%.
7	J. Malmsten, N. Zaninovic, Q. Zhan, M. Toschi, Z. Rosenwaks, J. Shan	Automatic prediction of embryo cell stages using artificial intelligence convolutional neural network	31,120 images of 100 mouse embryos from a public dataset. 661,060 images of 11,898 human embryos cultured in the TLM system	A stand-alone framework with a deep convolutional neural network (CNN) as the core for cell-stage image analysis on both datasets. The networks included Google's Inception architecture (V3) with and without transfer learning.	Cell division times were predicted accurately within five frames of the embryologist's annotation in 91% of the cell-stage transitions.
8	O. Barash, K. Ivani, L. Weckstein, M. Hinckley	High accuracy machine learning predictive model for embryo selection in IVF PGT cycles with single embryo transfers	1383 cycles (7120 embryos) of IVF PGT followed by 1108 SETs were included in the study (842 patients).	Combined predictions from multiple weak learners (GLM, random forest, gradient boosting, etc.) processed by Generalized Model Stacking produced a predictive performance of AUC	The probability of positive clinical outcome was calculated for each euploid embryo and ranged from 0.194 to 0.838 (baseline prediction - 0.645).
9	J.A. Gingold, N.H. Ng, J. McAuley, Z. Lipton, N. Desai	Predicting embryo morphokinetic annotations from time-lapse videos using convolutional neural networks	1309 embryos from 113 patients	A convolutional neural network	A per-frame ResNet model successfully classified frames into the appropriate developmental stage with 82% accuracy. After incorporating a late fusion model including the 14 surrounding frames, 84% accuracy was achieved, which improved to 87% following DP post processing (MAE 8.594, RMSE 24.334).
2017					
	K. Hunter Cohn, Q. Zhang, A. B. Copperman, P. Yurttas Beim	Leveraging artificial intelligence for more data-driven patient counseling after failed IVF-cycles	21,832 autologous IVF cycles from 13 centers	Extreme Gradient Boosting	AI models demonstrate that, while baseline patient metrics reflect overall prognosis, they do not provide additional information over multiple failed cycles. In contrast, retrieval metrics become increasingly important for refining counseling after failed cycles
European Society for Human Embryology and Reproduction 2018					
10	M. Meseguer Escriva, N. Zaninovic, F.G. Nogueira		223 Human embryo time lapse images.	Confusion matrices, ROC curves and Kappa Index	The AI's overall accuracy for prediction of blastocyst expansion (training

Table 1 (continued)

#	Authors	Title	Number of samples	Algorithms and/or architecture	Results
	Marcelo, O. Oliana, T. Wilkinson, L. Benham-Whyte, S. Lavery, C. Hickman, J.C. Rocha.	Using artificial intelligence (AI) and time-lapse to improve human blastocyst morphology evaluation			93.9% and validation 81.5%) and for prediction of ICM (training 93% and validation 78.8%) and trophectoderm (training 78.8% and validation 78.3%). The AI system was considerably more predictive of Expansion (AUC 0.888–0.956) compared to ICM (AUC 0.605–0.854) and trophectoderm (AUC 0.726–0.769).
11	T. Takeshima, S. Kuroda, M. Yamamoto, M. Murase, Y. Yumura, H. Sasaki, T. Hamagami.	Development of sperm searching system using artificial intelligence in assisted reproductive technology	8020 positive examples (spermatozoa) and 25,522 negative examples (non-spermatozoa)	Not described	Although the linear discrimination plane was learned from the histogram for both the learning data and the test data by the histograms of oriented gradient (HOG) feature extraction method, and the recall rate 0.99 was obtained, the false positive rate > 0.50 in the test data.
12	N. Correa, S. Brazal, D. García, M. Brassesco, R. Vassena.	Development and validation of an artificial intelligence based algorithm for the selection of an optimal stimulation protocols in IVF patients	Cohort database of 6952 first IVF cycles.	Algorithms tested included Random Forest (RF), and two types of black box algorithms such as Support Vector Machine (SVM, linear kernel), and Artificial Neural Network (ANN).	AI algorithm could predict with 81% accuracy the number of MII obtained by a patient only considering pre-treatment characteristics and comparing them to a large database of known cycles' characteristics, stimulation protocols and stimulation outcomes.
13	I. Sfontouris, A. Patelakis, G. Panitsa, S. Theodoratos, N. Raine-Fenning.	Comparison of machine learning models for the prediction of live birth following IVF treatment: an analysis of 463,669 cycles from a national database	463,669 cycles were included, of which 99,537 [(21.5% (95% CI: 21.4–21.6%)] resulted in a live birth.	Deep neural networks (DNN) was compared with random forest (RF), decision trees (DT) and Naive Bayes (NB) machine learning models.	The DNN model was associated with significantly higher accuracy, specificity, positive predictive value (PPV), positive and negative likelihood.
14	H. Matsubayashi, K. Kitaya, Y. Takaya, R. Nishiyama, K. Yamaguchi, N. Kim, K. Sakaguchi, M. Doshida, T. Takeuchi, T. Ishikawa	Identification of empty follicles or oocyte-containing follicles by ultrasound images using K-means method and principal component analysis assessing several parameters with artificial intelligence	Image data from 3 patients of 16 follicles (11 contained oocytes, 5 did not).	K-means method and principal component analysis	Oocytes-containing follicles were predicted 8/10 times and empty follicles were predicted 5/6 times.
15	Y. Miyagi, T. Habara, R. Hirata, N. Hayashi.	Feasibility of artificial intelligence for predicting live birth from a blastocyst image	80 Blastocysts resulting in live births and 80 resulting in abortion due to chromosomal abnormalities.	Logistic regression, naive Bayes, neural network, random forest, nearest neighbors and deep learning	The logistic regression with L2 regularization was the best machine learning method. The accuracy, sensitivity, specificity, positive predictive value and negative predictive value for predicting those in the abortion category were 0.619,

Table 1 (continued)

#	Authors	Title	Number of samples	Algorithms and/or architecture	Results
					0.600, 0.638, 0.626 and 0.629, respectively.
16	O.O. Barash, K.A. Ivani, S.P. Willmal, N. Huen, M.V. Homer, L.N. Weckstein. 2017	Model ensembling and synthetic features in PGS cycles with single embryo transfer	1029 cycles of IVF PGS treatment with single embryo transfer.	Generalized Model Stacking using generalized logistic regression, random forest, gradient boosting	Predictive performance of AUC = 0.8047
	K. Hunter Cohn, C. Glazner, L. Tan, E. Widra, M. Leondires, B. Miller, C. Benadiva, J. Nulsen, G. Letterie, A. Copperman, P. Yurttas Beim	Applying data-driven approaches to develop a novel ovarian reserve score that leverages underlying relationships between distinct markers of ovarian reserve	31,263 cycles from 26,929 patients.	Exploratory factor analysis (EFA) to group correlated variables that measure the same factor and structural equation modeling (SEM) was used to explore relationships between the factors uncovered by EFA.	EFA/SEM analyses revealed that ovarian reserve was defined by two distinct factors: one measured to a similar degree by BAFC and AMH (Factor 1) and the other by FSH and the LH to FSHratio (Factor 2).
	Abstract under press embargo	Using artificial intelligence to improve blastocyst morphology evaluation			

such as patient demographics and medical history, individualized hormone regimes, follicular growth patterns, endometrial window of implantation, pre-implantation genetic screening and diagnosis, and clinical pregnancy outcomes, are managed by incompatible systems.

Several "one-size fits all" aspects of infertility care could potentially benefit from a hyper-personalized AI or ML approach, including luteal phase progesterone supplementation and embryo transfer timing to the individual window of implantation, among others. One caveat exists though: we do not know what we do not know, and human experts have to "train" AIs, leading to potential amplification of bias, or completely over-looking certain types of data.

We do not yet know the feature or set of features that are most predictive of a successful IVF cycle. The most important variable(s) could be unknown-to-science and elucidated through use of AI. AI and ML methodologies seek to transcend the narrow focus on individual variables and uncover new epistemologies hidden in "big data."

Deep (artificial) neural networks (ANN) or convolutional neural networks (CNN) combine hardware and software to approximate the web of neurons in the human brain (AI/ML terms are defined in Table 2). The availability of high-tech central and graphics processing units (CPUs and GPUs), enormous data sets (i.e., big data), and developments in "machine learning" algorithms have lead to a stunning increase in the use of AI generally, across all fields, but it has been robustly adopted to medical imaging [7].

The learning part of "deep learning" is achieved during the training phase. Hundreds, or better yet millions, of 2D or 3D data points are fed into a model, so that future outputs can be predicted. Input values can be text, sound, signals, and most importantly for embryology, images. The fundamentals of deep learning methods for medical applications in general, image registration, anatomical/cell structure detection, tissue segmentation, computer-aided disease diagnosis, or prognosis have been reviewed extensively [8, 9].

Deep learning methods are most effective when large, unbiased datasets prevent "overfitting" (i.e., bias) of the model; however, massive datasets and massively parallel computing power are a bottleneck. Therefore, it is the large tech companies and institutions that are driving AI in-roads into healthcare. For example, IBM developed a computational model that predicts heart failure [10] and lent the Watson supercomputer to Memorial Sloan-Kettering for cancer diagnosis and selection of treatment. Watson for Genomics ingests approximately 10,000 scientific articles and 100 new clinical trials every month [11]. Stanford University reported a deep learning algorithm that predicts the safety of drug compounds and another to predict lung cancer type and patient survival, and Intel, who announced a competition to find an algorithm for early detection of lung cancer.

Table 2 Abbreviations and definitions

N/A	Algorithm	A set of defined step-by-step instructions. Can be very simple or very complex.
AI	Artificial intelligence	Not well defined. Broadly described as making a machine behave in ways that would be called "intelligent" if seen by a human.
ANN	Artificial neural network	A highly abstracted and simplified model compared to the human brain, used in machine learning. A set of units receives input data, performs computations on them, and passes them to the next layer of units. The final layer represents the answer to the problem.
N/A	Black box	The calculations performed by some deep learning systems between input and output are not easy (and potentially impossible) for humans to understand.
CNN	Convolutional neural network	In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery.
CPU	Central processing unit	The part of a computer in which operations are controlled and executed.
DL	Deep learning	A specific sub-field of deep learning. It is a process by which a neural network becomes sensitive to progressively more abstract patterns. Hundreds of successive layers of data representations are learned automatically through exposure to training data.
EMR	Electronic medical record	An electronic record of health-related information on an individual that can be created, gathered, managed, and consulted by authorized clinicians and staff within one health care organization.
ESA	Embryo selection algorithm	Any number of morphokinetic parameters that have been linked to an embryo's viability are combined, for example; the appearance and disappearance of pronuclei and nuclei at each cell stage, the length of time between early cytokinesis and initiation of blastulation, reabsorption of fragments, direct cleavage of cells within embryos from one to three cells, and reverse cleavage.
GPU	Graphics processing unit	A specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device.
ML	Machine learning	Algorithms that find patterns in data without explicit instructions. ML is a single contributing entity for AI technology.
PGT-A	Pre-implantation testing-aneuploidy	A set of techniques used on the embryo prior to transfer to the mother's uterus with the aim of studying any possible chromosomal and/or genetic abnormalities.
PPV	Positive predictive value	The post-test probability of being affected after a positive test.
SL	Supervised learning	A type of machine learning where the algorithm compares its outputs with the correct outputs during training.
N/A	Test dataset	The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.
N/A	Training dataset	The sample of data used to fit the model.
		The actual dataset that we use to train the model (weights and biases in the case of Neural Network). The model sees and learns from this data.
TL	Transfer learning	A technique in machine learning where the algorithm learns one task, and build on that knowledge while learning a different, but related, task. Transfer learning is an alternative approach to help mitigate the large, manually annotated data sets needed for training an AI.
N/A	Validation dataset	The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Given an infinite number of data points, time, and computing power, *any* continuous function can be approximated and predicted—even one as complex as an IVF cycle. Learning vector quantization networks [12] combined with computing in the "cloud" seem to suggest that AI for IVF could be right around the corner. However, the first step to implementing an AI is access to "big data" and many clinics have not yet implemented robust electronic medical record (EMR) systems.

There have been several types of machine-learning architectures proposed [13–16] for the IVF lab and infertility care. The goal is to integrate thousands (thousands or millions) of data points from electronic medical records, doctor's notes, and medical images, to uncover hidden clues that can diagnose infertility and predict the best course of treatment. Abstract 12 (Table 1) demonstrated that an AI was 81% accurate at predicting the number of MII oocytes obtained from a stimulation cycle, by considering pre-treatment characteristics and comparing them to a large database of known cycles' characteristics, stimulation protocols, and stimulation outcomes.

Although many people are participating in AI projects, few in human reproduction and embryology have been trained and educated on minimum acceptance standards. Although this proposed initiative may add an extra step to the abstract submission process, such a quality control has considerable value to reviewers, authors, and readers.

Consensus technical criteria have been established for reporting: authentication of human cell lines [17] (definitive standard ASN-0002 specifies methodology for DNA extraction, STR profiling, data analyses, and more), antibodies [18] (peptide/protein target, antigen sequence, name of antibody, manufacturer, catalog number or name of source, species raised in, monoclonal or polyclonal, and dilution used), next-generation sequencing-based cancer testing [19] (classification, annotation, interpretation, and reporting conventions for somatic sequence variants). Machine learning reporting standards have been proposed for biomedical applications (Table 1, Luo et al.) [20]. For abstracts, along with background and objective, the authors of the opinion suggest that abstracts should be structured to include "data sources, performance metrics of the predictive model or models in both point estimates and confidence intervals, and conclusion including the practical value of the developed predictive model or models."

In our survey of abstracts presented during the 2018 annual congresses of ASRM and ESHRE, we noted wide variation in reporting conventions.

Training, validation, and test data sets

By convention, 70% of any available data set is typically allocated for training. The remaining 30% are equally partitioned and referred to as "validation" and "test" data sets. With less training data, parameter estimates have greater variance. With less testing data, the performance statistic will have greater variance. Broadly speaking, the ideal division of data is such that neither variance is too high. Each of these data set size choices are equally as important.

A wide range of sample sizes were reported (Table 1): 3 patients with 16 follicles, 118 embryos, 160 blastocysts, 223 embryo images, 303 embryo with an associated 386 images, 34 with 773 oocytes, 50,392 images from 10,148 embryos, 463,669 cycles, and so on. Nearly all abstracts were missing an explanation for the choice of the size of training, validation, and test data set size for readers to know if the study was sufficiently powered to answer the problem. In other words, the number of parameters examined in the majority of problems was absent, as well as an explanation for us to confidentially make sure that enough information was examined to specify all the network's connections. The general rule of thumb in computer science is "the more data the better" but is there a minimum acceptable size for a training data set? Learning curves show model performance as a function of the training sample size and can help to determine the sample size needed to train, but they are not the only or final answer [21].

Data types

Additionally, the data types were variably reported as number of images, oocytes, embryos, patients, and cycles. Of course, not all studies will use more than one type of data. However, abstract #4 (Table 1) reported number of embryos ($n = 118$), images ($n = 2000$), but did not report on the number of patients. As a field, we may benefit from early adoption of standard reporting conventions so that "apples to apples" comparisons can be made confidentially across future studies.

Architecture and algorithms

In some cases, very general descriptions were reported, for example "a convolutional neural network" or "an artificial neural network (ANN) architecture associated with a genetic algorithm." In other abstracts, the architecture was fully described, for example "a stand-alone framework with a deep convolutional neural network (CNN) as the core for cell-stage image analysis on both datasets. The networks included Google's Inception architecture (V3) with and without transfer learning." In the machine learning abstracts, often no explanation was provided for why certain algorithms were selected, to help the reader determine the appropriateness of the tests or the comparisons.

Evaluating results

Some questions to consider in evaluating results are

- Were enough data points used to achieve a desired level of performance?
- Did the training set achieve a sufficient estimate of model performance?
- Was enough data analyzed to demonstrate that one model is better than another?
- Was an inflection point reached, where providing more training data to the system no longer yields improved accuracy?

Automated embryo selection by AI

Embryo development and selection is the natural starting point for the application of AI in the IVF lab, due to the availability of high-quality image data and importance of embryo selection [22] to success of an IVF cycle.

Embryo selection is based on subjective developmental and morphological characteristics. The necessary and sufficient quality or set of qualities to judge an embryo on has been explored ad infinitum; thickness of zona, granularity of cytoplasm, roundness of oolema, color of cytoplasm, multinucleation, number of blastomeres, degree of fragmentation, size of blastocoel, timing of cleavage and morphokinetics, and more.

Numerous experimental methods have been developed to parse the implantation potential and live birth rate of IVF embryos from images or video, including morphometric analysis by time-lapse imaging [23–26], mathematical and statistical tools [27, 28], and computer-assisted scoring [29, 30]. In 2008, the first commercially available time-lapse system (TLS) was sold for use in human in vitro fertilization (IVF) (Primo Vision™, Vitrolife, Göteborg, Sweden) when it was introduced at the European Society of Human Reproduction meeting. Incubator integrated TLSs currently available include the FDA-approved EmbryoScope®, Vitrolife; Miri®, ESCO (Egaa, Denmark); Geri, Genea Biomedx (Sydney, Australia).

Advancements in time-lapse imaging technologies have led to the development of embryo selection algorithms (ESAs). However, none of the proposed ESAs have surpassed a positive predictive value (PPV) of 45% (~ 38–44%) in selecting good quality embryos with relatively poor clinical outcomes [24, 31].

The FDA-approved Eeva Test [32] is one such ESA, driven by an Xtend algorithm (a standard multi-dimensional, static algorithm). Eeva uses time-lapse imaging (videos of embryo development) to predict which embryo has the best chance of progressing to a blastocyst, based not only on its appearance but when it hits certain developmental milestones (cell division timings P2 (time between first and second mitosis) and P3 (time between second and third mitosis).

Despite multiple lines of evidence for embryo selection, several excellent reviews [33–35] demonstrate that there is insufficient evidence in live birth, miscarriage, stillbirth, or clinical pregnancy to choose between TLS and conventional incubation.

AI has been shown to be better suited to some subjective tasks than trained embryologists. In one recent study, 482 seven-day-old bovine embryos were used to train an artificial intelligence system [36]. This analysis identified 36 assessment variables, 24 of which formed the input of the artificial network architecture. Overall, the artificial intelligence system had a 76% accuracy rate.

In human infertility, AI start-ups and products are already claiming significant results. For example, detection of chromosome 21 aneuploidy (Life Whisperer and Ovation Fertility), and CooperGenomicsSM PGTaiSM technology platform that improves the accuracy of PGT-A calling, interpretation and reporting while "removing subjectivity" in interpretation, and avoiding reporting transcription errors. Companies have no mandate to publish their results for peer-review, so it is difficult to evaluate any of these claims.

In 2017, we reported that the simultaneous evaluation of patients' entire embryo cohort at a single time-point could improve the identification of embryos with the highest potential to form a blastocyst [37]. We observed a higher overall blastocyst formation prediction rate with the cohort selection method (CM), compared to standard morphology grading by expert embryologists and time-lapse scoring (95% vs. 86% and 89%; $p = 0.06$ and $p = 0.21$, respectively).

Similarly, predicting high-quality blastocyst formation (expanded blastocysts with good inner cell mass and trophectoderm) was higher with the cohort method, compared to both standard morphology grading and time-lapse scoring (75% vs. 66% and 67%; $p = 0.21$ and 0.33, respectively). The prediction rate for overall and high-quality blastocyst formation was similar between standard morphology and time-lapse scoring. This small study was limited by a single time point, and retrospective data collection from a single, high-performing center.

To this end, we trained GoogleNet Inception v3 CNN architecture, replacing the final classification layer with an annotated dataset of embryo images captured with an EmbryoScope (submitted, under review 2018). This allowed us to prospectively interrogate data captured at multiple centers, using multiple time points to improve statistical significance, automatically detect patterns in image data, and utilize the uncovered patterns to identify the top quality embryo within a patient's cohort.

Further, we used the same architecture to report automated sperm morphology testing (Table 1, abstract 6) with transfer learning after replacing the final classification layer and retraining it with a dataset of sperm images.

Several other groups have also reported using AI for subjective problems, such as blastocyst morphology evaluation (Table 1, abstracts 2 and 10), and embryo selection (Table 1, abstracts 3, 5, and 8.)

Which brings us to our last point: true AI services essentially program themselves, and they do it in ways we do not fully understand. This is referred to in computer science as the "BlackBox" problem. Even the engineers who build AIs cannot fully explain their behavior. The potential for unintentional bias is enormous.

Notable examples of bias in AI applications include Google's AI algorithm labeling images of black people as gorillas, image searches for "CEO" returned only pictures of

white men with fewer ads for high-paying executive jobs displayed to women. A LinkedIn advertising program showed a preference for male names in searches, and there are many more examples. Private companies are racing to provide their products and services directly to the market, often without peer review. This is in contrast to academic AI projects, which have undergone extensive peer review and have (often) provided the source code itself to the reviewers for examination.

AI for embryo selection is being positioned on the market as a tool to prescreen for and identify viable embryos with a low likelihood of genetic defects *before* proceeding to PGT. This has the potential to result in significant cost savings for couples with many more embryos available to test than the standard "8 embryos" that are included in a typical PGT package. The availability of more than 8 good quality blastocysts is common with egg donor recipient cycles, but these are the cycles less likely to need PGT in the first place. The potential to discard large numbers of perfectly normal embryos with normal pregnancy potential in autologous IVF cycles is a significant concern. In this approach, AI for embryo selection seems to be predicated on PGT-A, the clinical utility of which is, in and of itself, a scalding hot topic [38–40].

A more pragmatic approach for the IVF laboratory who is interested in incorporating AI into their workflow may be to use it as a QC tool, for example, after embryo thawing, or to monitor embryo culture systems as a whole throughout the year.

With so much at stake in choosing embryos capable of producing a healthy child, it is our opinion that fully transparent, peer-reviewed projects should be prioritized over AI services where the source code, training data set, and validation data sets have not been examined closely for potential bias with respect to culture methodologies, patient populations, disease progressions, stimulation protocols, or dozens of other variables.

Conclusion

A clear cause cannot be identified in the majority of IVF failures [41]. Many studies are performed every year focused on individual variables. The field's focus has shifted from embryo selection, "freeze all" policies that allow for the uterine environment to be synchronized with embryo's development, to pre-implantation genetic testing. It is clear now that we are at the dawn of another step-change: using AI and ML to integrate all IVF cycle data, from the clinical unit that treats the primary infertility, to ovarian response to stimulation, to the IVF laboratory unit's embryo selection criteria, to the molecular diagnostic information for both PGT embryos and uterine receptivity, to obstetric and gynecological outcomes to achieve better outcomes for patients.

AI technology may become routine in clinical IVF settings within the next 5 years. The sweeping speed and promise is reminiscent of the application of PGT-A as an adjuvant treatment in IVF, which to this day remains controversial, because the weight of evidence in support of improving delivery rates is considered questionable [42].

If AI can differentiate "normal" embryos from those that are chromosomally abnormal with higher accuracy than other types of pre-implantation genetic testing, it promises to reduce costs, miscarriage, and stillbirth rate. Pragmatically, this technology could be valuable for patients, who do not wish to subject their embryos to biopsy, cannot afford PGT-A, and so on. An immediate and obvious application of AI in any IVF laboratory would be to adapt it for routine quality control and monitoring of culture systems.

We hope this perspective article stimulates a larger discussion toward supporting and progressing AI and ML for IVF. A committee opinion is needed to consider standardized reporting and minimal acceptance criteria for AI and ML studies, and the use of fully transparent peer-reviewed AI services, where the source code, training data, and validation data have been examined for potential bias.

References

1. Scientific Congress Supplement: Oral and Poster Session Abstracts. Fertil Steril. 2018;110(4):Supplement e1–e468
2. Scientific Congress Supplement: Oral and Poster Session Abstracts. Fertil Steril. 2017;108(3):Supplement e1–e422
3. Abstracts of the 33rd Annual Meeting of the European Society of Human Reproduction and Embryology. Hum Reprod. 2017;32(Supplemental 1):i1–i539
4. Abstracts of the 34rd Annual Meeting of the European Society of Human Reproduction and Embryology. Hum Reprod. 2018;33(Supplemental 1):i1–i541
5. European IVF-Monitoring Consortium (EIM) for the European Society of Human Reproduction and Embryology (ESHRE), Calhaz-Jorge C, de Geyter C, Kupka MS, de Mouzon J, Erb K, et al. Assisted reproductive technology in Europe, 2012: results generated from European registers by ESHRE. Hum Reprod. 2016;31(8):1638–52.
6. Kaufmann SJ, Eastaugh JL, Snowden S, Smye SW, Sharma V. The application of neural networks in predicting the outcome of in-vitro fertilization. Hum Reprod. 1997;12(7):1454–7.
7. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. 2016;35(5):1285–98.
8. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overview. Korean J Radiol. 2017;18(4):570–84.
9. Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng. 2017;19:221–48.

10. Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart WF. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. Circ Cardiovasc Qual Outcomes. 2016;9(6):649–58.
11. Bringing Precision Medicine to Community Oncologists. Cancer Discov. 2017;7(1):6–7
12. Siristatidis C, Vogiatzi P, Pouliakis A, Trivella M, Papantoniou N, Bettocchi S. Predicting IVF outcome: a proposed web-based system using artificial intelligence. In Vivo. 2016;30(4):507–12.
13. Meseguer M, Kruhne U, Laursen S. Full in vitro fertilization laboratory mechanization: toward robotic assisted reproduction? Fertil Steril. 2012;97(6):1277–86.
14. Siristatidis CS, Chrelias C, Pouliakis A, Katsimanis E, Kassanos D. Artificial neural networks in gynaecological diseases: current and potential future applications. Med Sci Monit. 2010;16(10):RA231–6.
15. Siristatidis C, Pouliakis A, Chrelias C, Kassanos D. Artificial intelligence in IVF: a need. Syst Biol Reprod Med. 2011;57(4):179–85.
16. Milewski R, Milewska AJ, Więsak T, Morgan A. Comparison of Artificial Neural Networks and Logistic Regression Analysis in Pregnancy Prediction Using the In Vitro Fertilization Treatment. Stud Logic Grammar Rhetoric. 2013;35(48):39–48.
17. Almeida JL, Cole KD, Plant AL. Standards for cell line authentication and beyond. PLoS Biol. 2016;14(6):e1002476.
18. Helsby MA, Fenn JR, Chalmers AD. Reporting research antibody use: how to increase experimental reproducibility. F1000Res. 2013;2:153.
19. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and guidelines for the interpretation and reporting of sequence variants in Cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017;19(1):4–23.
20. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12):e323.
21. Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. Anal Chim Acta. 2013;760:25–33.
22. Capalbo A, Rienzi L, Cimadomo D, Maggiulli R, Elliott T, Wright G, et al. Correlation between standard blastocyst morphology, euploidy and implantation: an observational study in two centers involving 956 screened blastocysts. Hum Reprod. 2014;29(6):1173–81.
23. Wong CC, Loewke KE, Bossert NL, Behr B, de Jonge CJ, Baer TM, et al. Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. Nat Biotechnol. 2010;28(10):1115–21.
24. 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. Fertil Steril. 2013;100(2):412–9 e5.
25. Kirkegaard K, Agerholm IE, Ingerslev HJ. Time-lapse monitoring as a tool for clinical embryo assessment. Hum Reprod. 2012;27(5):1277–85.
26. Rubio I, Galán A, Larreategui Z, Ayerdi F, Bellver J, Herrero J, et al. Clinical validation of embryo culture and selection by morphokinetic analysis: a randomized, controlled trial of the EmbryoScope. Fertil Steril. 2014;102(5):1287–1294 e5.
27. Cicconet M, Gutwein M, Gunsalus KC, Geiger D. Label free cell-tracking and division detection based on 2D time-lapse images for lineage analysis of early embryo development. Comput Biol Med. 2014;51:24–34.
28. Basile N, Vime P, Florensa M, Aparicio Ruiz B, García Velasco JA, Remohí J, et al. The use of morphokinetics as a predictor of implantation: a multicentric study to define and validate an algorithm for embryo selection. Hum Reprod. 2015;30(2):276–83.
29. Tian Y, Yin YB, Duan FQ, Wang WZ, Wang W, Zhou MQ. Automatic blastomere recognition from a single embryo image. Comput Math Methods Med. 2014;2014:628312.
30. Santos Filho E, et al. A method for semi-automatic grading of human blastocyst microscope images. Hum Reprod. 2012;27(9):2641–8.
31. 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(3):613–21.
32. Diamond MP, Suraj V, Behnke EJ, Yang X, Angle MJ, Lambe-Steinmiller JC, et al. Using the Eeva test adjunctively to traditional day 3 morphology is informative for consistent embryo assessment within a panel of embryologists with diverse experience. J Assist Reprod Genet. 2015;32(1):61–8.
33. Armstrong S, Arroll N, Cree LM, Jordan V, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. Cochrane Database Syst Rev. 2015;2:CD011320.
34. Armstrong S, Bhide P, Jordan V, Pacey A, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. Cochrane Database Syst Rev. 2018;5:CD011320.
35. Chen M, Wei S, Hu J, Yuan J, Liu F. Does time-lapse imaging have favorable results for embryo incubation and selection compared with conventional methods in clinical in vitro fertilization? A meta-analysis and systematic review of randomized controlled trials. PLoS One. 2017;12(6):e0178720.
36. Rocha JC, Passalia FJ, Matos FD, Takahashi MB, Ciniciato DS, Maserati MP, et al. A method based on artificial intelligence to fully automatize the evaluation of bovine blastocyst images. Sci Rep. 2017;7(1):7659.
37. Dimitriadis I, Christou G, Dickinson K, McLellan S, Brock M, Souter I, et al. Cohort embryo selection (CES): a quick and simple method for selecting cleavage stage embryos that will become high quality blastocysts (HQB). Fertil Steril. 2017;108(3):e162–3.
38. Gleicher N, Kushnir VA, Barad DH. How PGS/PGT-A laboratories succeeded in losing all credibility. Reprod BioMed Online. 2018;37(2):242–5.
39. Grati FR, Gallazzi G, Branca L, Maggi F, Simoni G, Yaron Y. Response: how PGS/PGT-A laboratories succeeded in losing all credibility. Reprod BioMed Online. 2018;37(2):246.
40. Munne S, et al. Response: how PGS/PGT-a laboratories succeeded in losing all credibility. Reprod BioMed Online. 2018;37(2):247–9.
41. Penzias AS. Recurrent IVF failure: other factors. Fertil Steril. 2012;97(5):1033–8.
42. Verpoest W, Staessen C, Bossuyt PM, Goossens V, Altarescu G, Bonduelle M, et al. Preimplantation genetic testing for aneuploidy by microarray analysis of polar bodies in advanced maternal age: a randomized clinical trial. Hum Reprod. 2018;33(9):1767–76.