Journal of Assisted Reproduction and Genetics (2021) 38:1641–1646
https://doi.org/10.1007/s10815-021-02198-x

ASSISTED REPRODUCTION TECHNOLOGIES

# Deep learning early warning system for embryo culture conditions and embryologist performance in the ART laboratory

Charles L. Bormann[1] · Carol Lynn Curchoe[2] · Prudhvi Thirumalaraju[3] · Manoj K. Kanakasabapathy[3] · Raghav Gupta[3] · Rohan Pooniwala[3] · Hemanth Kandula[3] · Irene Souter[1] · Irene Dimitriadis[1] · Hadi Shafiee[3]

Received: 10 September 2020 / Accepted: 18 April 2021 / Published online: 27 April 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021, corrected publication 2021

## Abstract

Staff competency is a crucial component of the in vitro fertilization (IVF) laboratory quality management system because it impacts clinical outcomes and informs the key performance indicators (KPIs) used to continuously monitor and assess culture conditions. Contemporary quality control and assurance in the IVF lab can be automated (collect, store, retrieve, and analyze), to elevate quality control and assurance beyond the cursory monthly review. Here we demonstrate that statistical KPI monitoring systems for individual embryologist performance and culture conditions can be detected by artificial intelligence systems to provide systemic, *early* detection of adverse outcomes, and identify clinically relevant shifts in pregnancy rates, providing critical validation for two statistical process controls proposed in the Vienna Consensus Document; intracytoplasmic sperm injection (ICSI) fertilization rate and day 3 embryo quality.

**Keywords** Clinical decision-making · Competency · Quality assurance · Proficiency · Embryo quality · Embryology · Laboratory quality management systems · Assisted reproductive technologies · Infertility, Artificial intelligence · AI · Convolutional neural network · CNN

## Introduction

The in vitro fertilization (IVF) laboratory director is responsible for identifying and monitoring key performance indicators of IVF laboratory success. It can be difficult to identify the source of problems when a program does not produce satisfactory pregnancy outcomes. All accredited laboratories must document continuous monitoring of quality control and assurance parameters, culture conditions, and competency assessments for staff [1]. More importantly, potential problems must be identified quickly to permit timely corrections—an effort that is hindered by the length of time to the clinical pregnancy test, manual data entry to spreadsheets, subjectivity, inaccurate recordings, and time-to analysis.

The clinical outcome of an IVF cycle is the gold standard of system quality, with ongoing pregnancy rates dependent on clinical KPIs [2, 3], culture systems [4], staff clinical decision-making, and technical competency to perform a wide range of procedures [5]. However, this approach has been criticized as not providing actionable insight soon enough, in the unfortunate event of a deleterious quality event. Alternate factors beyond the laboratory's control, such as the number of good quality embryos available, number of good quality embryos transferred, and number of embryos suitable for freezing [6], also demonstrate significant correlation to pregnancy rates.

The primary outcome used to analyze embryology staff proficiency in performing intracytoplasmic sperm injection (ICSI) is fertilization rate. This outcome is measured between 16 and 18 h after insemination, and provides little information about the quality of the resulting embryo. Fertilization checks and embryo quality assessments require manual examination, recording of status, and embryo developmental scores. These processes are labor-intensive and subjective.

✉ Carol Lynn Curchoe
cburton@fertilitylabsciences.com

[1] Massachusetts General Hospital Fertility Center, Obstetrics/Gynecology/Reproductive Endocrinology and Infertility, Boston, MA, USA

[2] Colorado Center for Reproductive Medicine, Newport Beach, CA, USA

[3] Division of Engineering in Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Artificial intelligence (AI) in health care has shown the most promise in diagnostics, especially image-based analysis where AI systems can process "big" data and information to help arrive at clinically useful conclusions and recommendations. AI systems have been significantly investigated in the past several years for a wide variety of assisted reproductive technology (ART) applications [7]. Notably, AI systems have been developed for embryo selection [8, 9] and IVF cycle outcome prediction [10, 11]. However, this advanced technology has not been yet been demonstrated as a tool for monitoring individual embryologist performance or for quality assurance in an ART laboratory.

The goals of this study were to validate the predictive power of a previously developed AI to (1) detect performance shifts of embryologists conducting ICSI, and (2) identify early warning indicators for embryo culture conditions.

Here we present the validation of an AI algorithm, which predicts in vitro human embryo developmental fate, to both monitor the performance of embryo culture systems and to evaluate individual embryologist's performance of ICSI. The AI-generated predictions were found to have a high association with pregnancy rates ($R^2$=0.9063) and low variation with individual embryologist performance when compared to other KPI techniques.

## Materials and methods

**Image capture and annotation** Data was collected at a single fertility center in Boston, MA, under an institutional review board approval (IRB#2017P001339). EmbryoScope videos were fragmented to extract the frames linked to specific time points using a custom python script, which made use of the OpenCV and Tesseract libraries. Machine-generated timestamps available on each frame of the video were used to identify the images associated with specific time points. All embryos used in the study were annotated using images from the fixed time-points by senior level embryologists with a minimum of 5 years of human IVF training. Out-of-focus images were included in the datasets and used for both testing and training. Only images of embryos that were completely non-discernible were removed from the study as part of the data cleaning procedure.

### Automated fertilization and blastocyst assessment

Using annotated data of 2366 embryos, we trained and validated our convolutional neural network (CNN) to categorize zygotes based on their number of pronuclei (Fig. 1). We then evaluated the ability of the CNN in classifying zygotes based on the number of pronuclei using a test set of 947 embryos. The accuracy of the algorithm in 2PN and non-2PN embryo classification using the 947 embryos test set was 93.1% with a confidence interval (CI) ranged from 91.3 to 94.6%. We also used this dataset of embryos to train a network using day 5 morphology embryos [12, 13]. We evaluated the accuracy of the CNN in classifying embryos based on morphology on day 5 using a test set of 742 embryos. The accuracy of the system in categorizing embryos into two classes of blastocysts and non-blastocysts was 90.2% (CI: 87.8 to 92.2%**).**

### Early developmental stage markers as predictors for KPI monitoring

A deep neural network (AI) [8] analyzed embryo images acquired at 70 h post-insemination and provided a score (KPI score) taking into account all embryos within a given group. A total of 876 embryos (Fig. 2) were cultured in 6 different lots of media (Media A-F; CSC-Complete, Irvine Scientific), under identical conditions at 37°C, 5% $O_2$, and 6.5% $CO_2$ with oil overlay (Ovoil, Vitrolife) over a 6-month time period. The percentage of 2 pronuclei (2PN) zygotes at the 4-cell stage on day 2, 8-cell, 6- to 10-cell, ≥ 7-cells, and those predicted to develop into high-quality blastocyst stages using an AI-based generated KPI on day 3 of embryo development, were compared with ongoing pregnancy rates using a regression analysis. The low threshold value for ongoing pregnancy rates at this hospital-based IVF practice is set at 50%.

### AI algorithm architecture and dataset

To analyze and provide a KPI score for the embryos, we designed a convolution neural network-based deep learning technique. Here we used Xception architecture, which is a combination of depth wise separable convolution layers with residual connections comprising of 36 convolutional layers forming the feature extraction. We pre-trained with 1.4 million images from ImageNet, which performed with a top-1 accuracy of 79% and top-5 accuracy of 94.5% across 1000 classes of ImageNet database. We trained, evaluated, and tested this network with cleavage stage embryo images of 2449 embryos categorized across five classes based on their clinical annotations and their developmental fate at 113 h post-insemination (hpi). At 113hpi, class 1 consisted of degenerated and arrested embryos; class 2 embryos were at the morula stage; class 3 embryos were early stage blastocysts with small blastocoel cavities, indistinguishable inner cell masses, and trophectoderm cells; class 4 embryos were expanded blastocysts that did not meet freezable quality criteria based on the practice guidelines (> 3CC), where 3 represents the degree of expansion (range 1–6) and C represents the quality of ICM and TE (range A–D), respectively; and class 5 are embryos that met freezing criteria included full to hatched blastocysts. These embryos have an A or B ICM and/or trophectoderm. A total of 2449 embryos were divided into a training dataset of 1190 images, validation set of 511 images, and test set of 748 non-overlapping images of cleavage stage

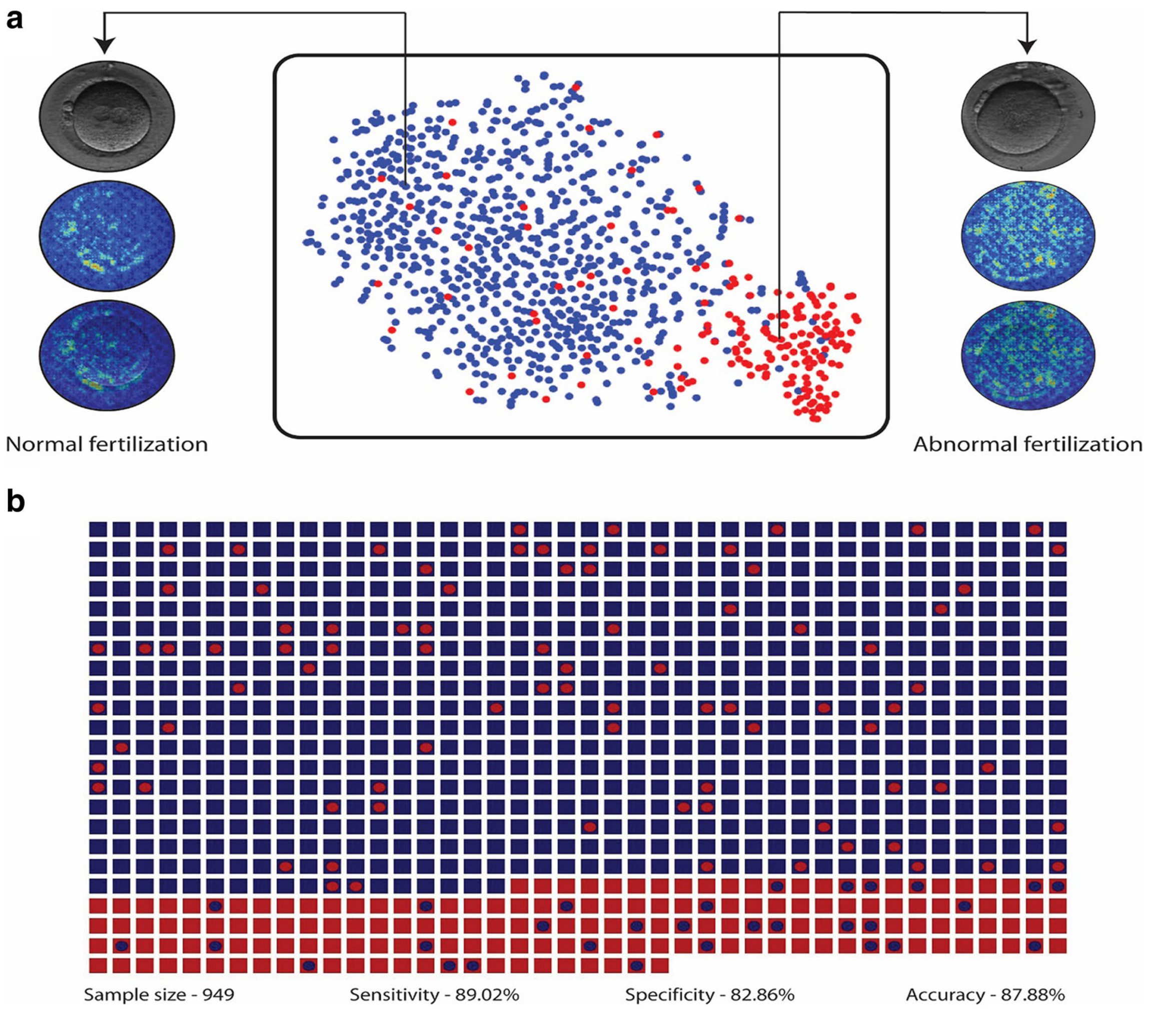

**Fig. 1** Fertilization assessment. **a** The t-SNE plot for the Xception model trained to classify abnormally fertilized embryos (non-fertilized, 3PN, 1PN etc. embryos) and normally fertilized embryos (2PN embryos). The saliency map of the two embryos provides an example of the features that network uses to classify embryos at the pronuclear stage. **b** The dot matrix plot illustrates the system's performance in evaluating embryos (*n*=947) from the test set of patients. The squares represent true labels and the circles within them represent the system's classification. Blue squares and circles represent normally fertilized embryos while red squares and circles represent abnormally fertilized embryos

embryos. Using this test of 748 embryos, the accuracy of the algorithm in predicting blastocyst development at 70 hpi was 71.87% [14]. We further used an independent test set of 876 embryos to generate KPI scores.

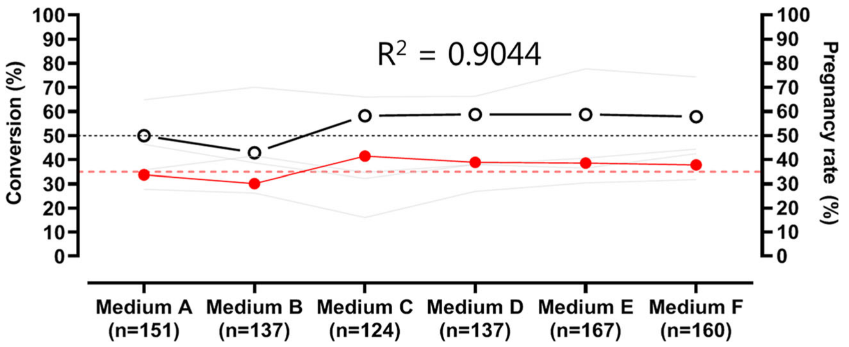

**Fig. 2** Early developmental stage markers as predictors for KPI monitoring. A deep neural network (AI) [8] analyzed embryo images acquired at 70 h post-insemination and provided a score (KPI score) taking into account all embryos within a given group. A total of 876 embryos were cultured in 6 different lots of media (Media A-F; CSC-Complete, Irvine Scientific) and under identical conditions at 37°C, 5% $O_2$, and 6.5% $CO_2$ with oil overlay (Ovoil, Vitrolife) over a 6-month period

### Training and implementation

All the layers of the network were trained, and the classification layer was replaced with a fully connected classification layer with random weights for class 5 embryo classification. Random rotation and flip argumentations were performed on the training set across all classes during training to improve the generalizability. The algorithm was implemented in python 3.7 using Keras, OpenCV libraries and trained on GTX 1080ti GPUs for 200 epochs using SGD optimizer with a batch size of 64 and a learning rate of 0.00075. The network generates 5 class confidence probabilities, which were used as a KPI measure.

### Automated quality assessment of individual embryologists performing ICSI using AI

The second goal of the study was to determine whether an automated AI system could be used to accurately monitor the performance of embryology staff performing ICSI. Over the course of 6 months, the developmental outcomes from 7 embryologists performing ICSI were tracked using standard manual morphologic measurements and KPI calculations. The first KPI analyzed was fertilization rate. Fertilization is evaluated 16–18h after ICSI to determine the presence or absence of pronuclei (PN). Normal fertilization is defined as formation of two distinct pronuclei. The next KPI analyzed is the blastocyst development rate. This KPI calculates the percentage of 2PN embryos that develop to the blastocyst stage within 5 days of culture (class 3–5 descriptions). The last KPI manually calculated is the high-quality blastocyst (HQB) conversion rate. This measurement calculates the percentage of 2PN embryos that develop to a freezable stage by day 5 of development (see class 5 description). All three of these KPIs require manual morphological assessments at fixed developmental time points under high power magnification. These 3 manually assessed KPIs were compared against an AI system developed to automate the morphological assessment of embryos at the same stages of development. The rates of fertilization, blastocyst development, and high-quality blastocyst (HQB) development were compared in a total of 947 embryos, divided between 7 embryologists. To evaluate the difference between the two analysis methods, we performed a Wilcoxon matched-pairs signed rank test and a coefficient of variation (%CV) analysis.

## Results

### Developmental fate of ICSI-derived embryos

The Wilcoxon tests revealed that the two approaches performed with negligible differences ($P$>0.05) for all three rate estimations (fertilization, blastocysts, and HQB). Figure 3 shows the medians of difference for estimations of fertilization, blastocysts, and HQB which were −1.3% ($P$>0.31), 1.8% ($P$>0.09), and −3.6% ($P$>0.18), respectively. The %CV estimations also showed that the difference between manual and AI-generated estimations for each embryologist in all three rates was low. The median of %CV between the two approaches in measuring the rates of fertilization, blastocysts, and HQB was 1.9%, 3.4%, and 10.9%, respectively.

### Early developmental stage markers as predictors for KPI monitoring

The AI-based-generated KPI for predicting high-quality blastocyst formation had the highest association with ongoing pregnancy rates ($R^2$=0.9063). This was the only cleavage stage KPI examined that was able to detect changes in our embryo culture environment that resulted in the pregnancy rates dropping below the threshold of 50% (Table 1).

## Discussion

Without question, laboratory quality control and assurance must be performed routinely in an IVF lab, with the goal of maintaining optimal culture conditions that leads to a healthy, live-born baby [15, 16]. While the embryologist's role in achieving and contributing to quality through safety in the assisted reproduction lab is well documented [16], appropriate levels of monitoring, what to monitor, and the best ways to monitor it are unclear and far from standardized. Laboratory staff spend countless hours monitoring (sometimes multiple times a day) and ensuring that the laboratory staff, equipment, and environment remain within the parameters of the lab's QC program.

Many IVF labs, at a minimum, record the batch number of all culture media, disposables, and laboratory ware used for a particular patient and check incubators every day with regard to temperature, humidity, and atmospheric conditions [17]. Consumables and physical culture environment are undeniably important; however, staff competency is also a crucial component of the IVF laboratory's quality management system. The "human factor" among physicians [2] can significantly affect ongoing clinical pregnancy rates. Cirillo et al. (2020) examined operator effect in frozen embryo transfers and found that from worst to best operator, the odds ratio varied between 0.84 and 1.13. The odds of success with the worst operator were almost 16% lower than the mean, and the odds of success for the best operator was 13% higher. Likewise, embryologists must be competent to make dozens of clinical decisions that can affect cycle outcomes, and be technically proficient in a wide range of procedures.

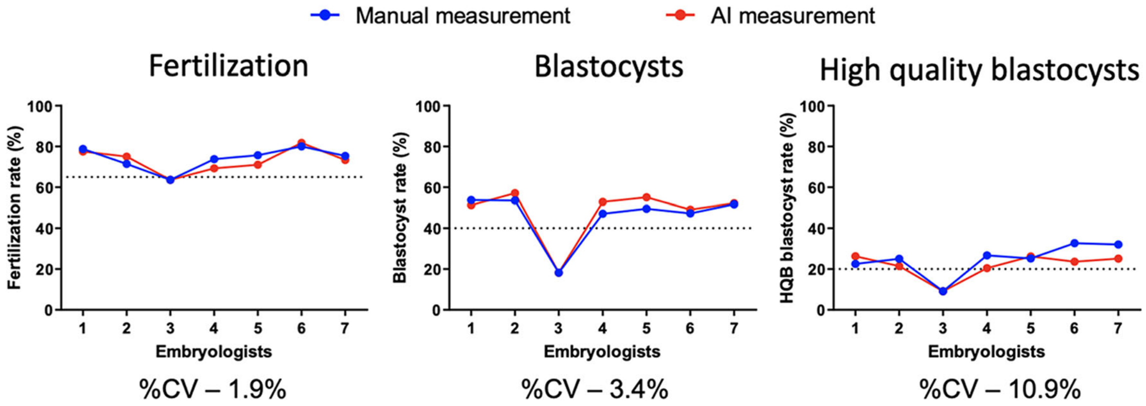

**Fig. 3** Comparison results. The Wilcoxon tests revealed that the two approaches in performed with negligible differences ($P$>0.05) overall for all three rate estimations (fertilization, blastocysts, HQB)

Certain IVF key performance indicators were formally defined in the Vienna Consensus document published in 2017 [18]. Reference indicators, such as oocyte maturity rate, show how the patient clinical side of the practice is performing. Laboratory performance indicators fall into two categories—those with a minimal threshold value (sperm motility post prep) and ones with a simple threshold limit (such as IVF polyspermy rate). The Vienna Consensus document defines blastocyst development rate as a KPI, but does not support use of a high quality or "usable" blastocyst rate. Blastocyst utilization rate was considered for the Vienna Consensus, but rejected as a KPI. Instead, the blastocyst formation rate, with no consideration of blastocyst-stage or blastocyst quality, was used.

The effectiveness of using these KPIs for detecting clinically relevant shifts following changes in laboratory processes was unverified, until a recent report by Hammond et al. [19] demonstrated the utility of day 5 blastocyst rate in a statistical KPI monitoring system to provide systematic, early detection of adverse outcomes in ART laboratories. Hammond et al. further extended and defined the day 5 blastocyst rate as a "usable" blastocyst metric, i.e., not just blastocyst development respective of quality but also good quality, usable blastocysts. Here we present data to support a KPI of "high-quality blastocyst (HQB) conversion rate." This measurement calculates the percentage of 2PN embryos that develop to a freezable stage by day 5 of development. The AI-based-generated KPI for predicting high-quality blastocyst formation had the highest association with ongoing pregnancy rates.

An important aspect of quality assurance data analysis is identifying statistical process controls that will provide meaningful insight into laboratory functioning at the earliest possible time point. Clinical pregnancy rates, although the gold standard, are not the only outcome worth measuring, and those data are only available after the "two-week wait" with clinical fetal heart rate confirmation coming much later. Factors such as multiple pregnancies, ovarian hyperstimulation, patient satisfaction, and the proper evaluation of laboratory and clinical protocols are also important metrics [20]. The timely discovery of a struggling technologist or bad lot of consumables, followed by immediate corrective actions through effective quality management practices, is the ambition of a quality control and assurance program. Table IV of the Vienna Consensus document indicates that ICSI fertilization rate and day 3 embryo development rate (defined as number of embryos with 8 or more cells on day 3 are among the key performance indicators to be recorded and analyzed [18].

Here we demonstrate for the first time, the power of using AI predictions in monitoring the performance of individual embryologist technical competency and early embryo

**Table 1** Early developmental stage markers as predictors for KPI monitoring

| Key performance indicator | Medium A ($n$=151) | Medium B ($n$=137) | Medium C ($n$=124) | Medium D ($n$=137) | Medium E ($n$= 167) | Medium F ($n$=160) | $R^2$ |
|---|---|---|---|---|---|---|---|
| Day 2: % 4-cell | 35.8 | 41.6 | 34.7 | 38.0 | 36.5 | 42.5 | 0.01144 |
| Day 3: % 8-cell | 27.8 | 26.3 | 16.1 | 27.0 | 30.5 | 31.9 | 0.01144 |
| Day 3: % 6–10 cell | 56.4 | 38.7 | 32.3 | 38.0 | 40.7 | 44.4 | 0.0415 |
| Day 3: % ≥ 7-cell | 64.9 | 70.1 | 66.1 | 66.4 | 77.8 | 74.4 | 0.0557 |
| Day 3: % AI-generated KPI | 33.8 | 30.0 | 41.5 | 38.9 | 38.6 | 37.8 | 0.9063 |
| % ongoing pregnancy rate | 50 | 42.9 | 58.3 | 58.8 | 58.8 | 57.9 | |

developmental stage markers as a predictor for the embryo culture environment. Furthermore, we demonstrated the link between quality assurance performance and patient outcomes.

This study is the first to describe that artificial intelligence could be used to automate the monitoring of individual embryologists performing ICSI in a clinical setting. Ideally, this remote monitoring will be performed through a laboratory information management system (LIMS) that can augment in real-time image features with competency assessments and many other types of patient-related clinical and laboratory KPI data [21]. An embryologist quality assurance software should automatically provide "red flags" when panic values are reached. The extremely low coefficient of variation between the manual and AI-based QA assessment methods demonstrates the high accuracy of the AI system.

A limitation of the approach presented here is the omission of clinical KPIs that are significantly associated with pregnancy rates, (age, AMH, and number of oocytes collected) and have proven to be useful analysis tools to combine with laboratory KPIs to predict the rates of clinical gestation [3]. Furthermore, the present study was limited in its analysis of other laboratory KPIs that can influence the fertilization rate, such as sperm quality, age of oocytes, oocyte maturity, or other variables.

In conclusion, this study demonstrates an alternative AI-driven method for monitoring two of KPIs noted in the Vienna consensus document in the IVF laboratory, without the need for subjective grading, manual recording, and analysis. Our work further validates the effectiveness of statistical process controls for detecting clinically relevant shifts in culture conditions and individual embryologist competency.

## References

1. Olofsson JI, Banker MR, Sjoblom LP. *Quality management systems for your in vitro fertilization clinic's laboratory: why bother?* J Hum Reprod Sci. 2013;**6**(1):3–8.
2. Cirillo F, Patrizio P, Baccini M, Morenghi E, Ronchetti C, Cafaro L, et al. *The human factor: does the operator performing the embryo transfer significantly impact the cycle outcome?* Hum Reprod. 2020;**35**(2):275–82.
3. Franco JG Jr, et al. *Key performance indicators score (KPIs-score) based on clinical and laboratorial parameters can establish benchmarks for internal quality control in an ART program*. JBRA Assist Reprod. 2017;**21**(2):61–6.
4. Castillo CM, et al. *The impact of selected embryo culture conditions on ART treatment cycle outcomes: a UK national study*. Hum Reprod Open. 2020;**2020**(1):hoz031.
5. Matson PL. *Internal quality control and external quality assurance in the IVF laboratory*. Hum Reprod. 1998;**13**(**Suppl 4**):156–65.
6. Strandell A, Bergh C, Lundin K. *Selection of patients suitable for one-embryo transfer may reduce the rate of multiple births by half without impairment of overall birth rates*. Hum Reprod. 2000;**15**(12):2520–5.
7. 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;**36**(4):591–600.
8. 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 Chip. 2019;**19**(24):4139–45.
9. 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 Digit Med. 2019;**2**:21.
10. Vogiatzi P, Pouliakis A, Siristatidis C. *An artificial neural network for the prediction of assisted reproduction outcome*. J Assist Reprod Genet. 2019;**36**(7):1441–8.
11. 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*. Hum Reprod. 2019;**34**(6):1011–8.
12. Thirumalaraju P, Kanakasabapathy MK, Bormann CL, Gupta R, Pooniwala R Kandula H, Souter I, Dimitriadis I, Shafiee H. Evaluation of deep convolutional neural networks in classifying orphological quality. https://www.arxiv.org/abs/2005.10912
13. Behr B, Wang H. *Effects of culture conditions on IVF outcome*. Eur J Obstet Gynecol Reprod Biol. 2004;115(Suppl 1):S72–6. https://doi.org/10.1016/j.ejogrb.2004.01.016.
14. Kanakasabapathy, M.K., et al. Deep learning mediated single time-point image-based prediction of embryo developmental outcome at the cleavage stage. arXiv:2006.08346v1 [q-bio.TO]
15. Smith GD, Takayama S, Swain JE. *Rethinking in vitro embryo culture: new developments in culture platforms and potential to improve assisted reproductive technologies*. Biol Reprod. 2012;**86**(3):62.
16. Go KJ. '*By the work, one knows the workman': the practice and profession of the embryologist and its translation to quality in the embryology laboratory*. Reprod BioMed Online. 2015;**31**(4):449–58.
17. Wikland M, Sjoblom C. *The application of quality systems in ART programs*. Mol Cell Endocrinol. 2000;**166**(1):3–7.
18. Embryology, E.S.I.G.o. and c.b.g.i. Alpha Scientists in Reproductive Medicine. Electronic address, The Vienna consensus: report of an expert meeting on the development of ART laboratory performance indicators. Reprod BioMed Online. 2017;**35**(5):494–510.
19. Hammond ER, Morbeck DE. *Tracking quality: can embryology key performance indicators be used to identify clinically relevant shifts in pregnancy rate?* Hum Reprod. 2019;**34**(1):37–43.
20. Alper MM, Brinsden PR, Fischer R, Wikland M. *Is your IVF programme good?* Hum Reprod. 2002;**17**(1):8–10.
21. Curchoe CL. Smartphone applications for reproduction: from rigorously validated and clinically useful to potentially harmful. EMJ Repro Health. 2020;6(1):85–91.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.