COMPREHENSIVE AI-BASED SYSTEM FOR KPI ASSESSMENT IN IVF LABORATORIES: DEVELOPMENT AND IMPLEMENTATION

***SERGEI SERGEEV**, PhD,[1] **IULLIA DIAKOVA,**[2] **LASHA NADIRASHVILI**[3]*
[1] Georgian-German Reproduction Center, Tbilisi, Georgia; IVF and Genetics Center, Moscow, Russia
[2] ART-IVF Clinic, Moscow, Russia
[3] Georgian-German Reproduction Center, Tbilisi, Georgia

SUMMARY

Background: In vitro fertilization (IVF) success heavily relies on laboratory Key Performance Indicator (KPI) evaluation systems, but traditional analytical methods often fail to provide deep, actionable insights into KPI establishment and measurement.

Objective: This study aimed to develop a comprehensive AI-based KPI evaluation system for IVF laboratories, focusing on a novel Deep Neural Network (DNN) model for predicting clinical pregnancy rates (CPR).

Materials and Methods: We analyzed 3,888 IVF treatment protocols, utilizing consensus KPIs from guidelines, various analytical methods, such as descriptive statistics, time series analysis, machine learning algorithms, and a custom-developed DNN model.

Results: The DNN model demonstrated a high level of accuracy, with an AUC of 0.79 and a PRC of 0.69 in predicting CPR. Importantly, there was no significant difference ($p < 0.05$) between predicted and actual CPR, reaffirming the model's reliability. The model's performance was on par with PGT-A-tested embryo transfers and other commercial AI solutions in IVF, including time-laps systems. External validation in independent clinics yielded an AUC of 0.73, consistent with cross-validation reports from multiple clinics. The application of the DNN model in 4 clinics for quality assurance identified variations in individual staff performance, enabling targeted mentoring and quality improvement, further reinforcing the system's reliability and reproducibility.

Conclusion: The developed AI-based KPI assessment system is a significant leap forward for IVF analytics. It provides a comprehensive, accurate, and reproducible tool for internal quality assurance, external clinic audits, and individual staff competency assessment. By shifting the focus from traditional embryo selection to a deeper understanding of parameters influencing successful IVF outcomes, it opens the door to more personalized and effective infertility treatments, offering hope for the future of IVF.

Keywords: in vitro fertilization; quality assurance; artificial intelligence; deep neural networks; key performance indicators; clinical pregnancy prediction; predictive modeling.

INTRODUCTION

In vitro fertilization (IVF) is a complex and critical process in assisted reproductive technology. The success of it heavily relies on the expertise and performance of embryologists. Implementing a robust Key Performance Indicator (KPI) evaluation system is essential to ensure optimal outcomes and continual improvement. This study aims to develop a comprehensive KPI evaluation system for an IVF clinic's embryology laboratory using artificial intelligence (AI) algorithms, detailing the data analysis process, its importance, and the advantages of such an approach.

Historically, the field of IVF has relied on a limited set of analytical tools for assessing laboratory performance and patient outcomes. Conventional data analysis typically involves basic descriptive statistics (means, medians, standard deviations), simple success rate calculations, and the occasional use of basic inferential statistics. While these methods provide a fundamental understanding of performance, they often fail to offer deep, actionable insights into the complex dynamics of IVF processes. A robust, continuous quality control system is paramount in the high-stakes field of assisted reproduction. Our AI-based solutions can find a place in this system as independent, reproducible, and highly effective algorithms for analytics and data-driven decision-making approaches.

In recent decades, neural networks and machine learning (ML) models have become vital tools in various fields, including computer vision, natural language processing, recommender systems, medical diagnostics, and the field of IVF.[1] These models serve as the foundation for creating algorithms capable of extracting complex dependencies from data, making predictions, and making decisions based on these dependencies. With that AI-based approach, we can transform our descriptive retrospective analytics into predictive prospective research. In some cases, neural networks can identify a broader spectrum of associations than other statistical methods, thanks to their ability to recognize highly nonlinear associations among input parameters. Therefore, the critical point of our study was the establishment of a new approach to KPI analytics utilizing deep learning network (DNN) architecture to predict the treatment cycle result – clinical pregnancy achievement (CPR).

The importance of the laboratory stage in IVF is detailed in consensus resolutions for quality control and the assessment of KPIs, constituting an integral part of internal quality control (QC) according to international standards.[2] Nevertheless, determining such indicators to identify potential problem areas in laboratory work is not always straightforward, and clear recommendations on which KPIs need improvement for real increases in successful IVF protocol numbers are lacking.[3,4] KPIs provide a comprehensive view of the embryologist's performance at various stages of the IVF process, allowing for a nuanced understanding of success rates and potential areas for improvement. In this work, we utilized consensus KPIs as a tool for outcome prediction and laboratory performance measurement.

METHODS

To create the KPI assessment system, we selected the Vienna Consensus[2] and the Maribor Consensus[5] as reference quality indicators, with adjustments for determining the total number of good-quality blastocysts and individualized KPI calibration according to patient population data from the publication by Zacà et al.[6] and the ASPIRE guidelines.[7]

We used retrospective data of 3888 IVF treatment protocols with known outcomes in "The Georgian-German Reproduction Center," Tbilisi, Georgia, from January 2022 to January 2024 to develop a data set for ML training and 394 protocols for model testing. For validation, a PGT-A dataset of 1600 cycles was used. For external model validation, we used data from 2 independent ART centers in Russia. All protocols with missing data values were discarded from the study.

Patient informed consent for that study was unnecessary because only retrospective and fully de-identified data from embryo development has been used. It is entirely non-invasive for patients or their embryos (no medical intervention was performed on the subject, and no biological samples from the patient were collected to develop that model). The ESHRE recommendations and Gardner blastocyst grading system were used for embryo evaluation, in which "good blastocysts" were identified as Bl3BB and higher grades.

Python 3.11, Scikit-learn 1.4.2, and Sklearn 1.4 were used to implement machine learning models and statistical modeling in DataSpell 2024.1.3 IDE. The neural network model has been developed and executed in the GPU PyCharm 17.0.10 environment with the Tensorflow 2.15.0 and Keras library 2.14.0. DNN calibration was performed using CalibratedClassifierCV from sci-kit-learn, which applies logistic regression to align probabilities. A comparative analysis of prediction errors was conducted with area under the receiver operating characteristic curve (AUC), accuracy, F-1 score, specificity (actual negative rate), sensitivity (recall), precision (positive prediction value), precision-recall curve (PRC) and Matthews Correlation Coefficient (MCC).

Statistical analysis of individual KPIs was conducted using StatTech software version 3.0.6. Descriptive statistics were chosen based on data distribution: for customarily distributed quantitative indicators, mean (M) and standard deviation (SD) with 95% confidence intervals (CI) were used, while median (Me) and interquartile range (IQR: Q1-Q3) were employed for non-normally distributed data, as determined by the Shapiro-Wilk test. The direction and strength of correlation between two quantitative variables were evaluated using Spearman's rank correlation coefficient. A p-value < 0.05 was used as the significance threshold for statistical analysis. Comparison of groups based on quantitative indicators was performed using one-way analysis of variance (ANOVA) for normally distributed data or the Kruskal-Wallis test for non-normally distributed data, followed by post hoc comparisons when significant differences were detected.

RESULTS AND DISCUSSION

Data preparation

The first step in developing our KPI evaluation system was to collect and prepare the relevant data. From the protocols of treatment cycles, we selected those executed by the current team of embryologists present in the laboratory. We established clear criteria for excluding protocols, including incomplete data, information provision errors, and procedures performed by multiple embryologists. We chose protocols that satisfied the Vienna consensus criteria to prepare the dataset for individual staff KPI analysis. Our statistical analysis revealed that all parameters, except the patient's age, exhibited distributions different from usual (R-test), and according to the Dickey-Fuller criterion, their time series were non-stationary ($p > 0.05$). We analyzed the workload distribution in the laboratory among staff based on input (number of oocytes used in procedures) and output (number of obtained blastocysts) parameters. This analysis revealed no significant differences ($p > 0.05$) in pairwise comparisons between all embryologists. This approach to data preparation is crucial for the validity and reliability of our subsequent KPI evaluation and analysis, allowing us to draw meaningful Conclusion about the performance of our IVF laboratory and individual embryologists.

Descriptive statistical analysis

We conducted descriptive statistical analysis to gain insights into the overall performance and variability of the KPIs. The mean patient age was 34 (SD = 5.7, min 18, max 49) years; the number of follicles 16.3 (SD = 9.64); OCC number 13.04 (SD = 7.32); number of oocytes used for fertilization 11.19 (SD = 6.68); 2pN number 8.15 (SD = 5.37), number of cleavaged embryos 8.12 (SD = 5.36); total number of blastocyst D5-D6 3.33 (SD = 3.43); number of good quality blastocyst 3.14 (SD = 3.67); fertilization rate 0.74 (SD = 0.31); cleavage rate 0.99 (SD = 0.4); reasonable blastocyst rate 0.40 (SD = 0.35); oocyte retrieval rate 0.78 (SD = 0.5).

The KPIs we analyzed provide a comprehensive view of embryologists' performance at various stages of the IVF process, allowing for a nuanced understanding of success rates and potential areas for improvement. Our quarterly analysis for 2022-2023 focused on input parameters for the laboratory, with particular attention to the main criterion of properly selected stimulation - the MII rate. This analysis revealed stability in the obtained mature oocytes, with a median of 0.87 and an interquartile range (Q1-Q3) of 0.855-0.893. These values consistently surpassed the Maribor consensus competency threshold of≥74% and a target value of ≥90%.[5] Notably, we observed invariance in these rates across the studied years. Furthermore, we conducted a comparative analysis of the main clinic KPIs between IVF and ICSI procedures. Our statistical tests revealed no significant differences in several key areas: the fertilization rate (U statistic = 4687.0, $p = 0.2945$), blastocyst formation rate (U statistic = 6096.5, $p = 0.0528$), good-quality blastocyst formation rate (U statistic = 4920.5, $p = 0.9031$), and availability of MII oocytes (U statistic = 5404.0, $p = 0.8479$). These findings suggest consistency in performance across different fertilization techniques employed in our clinic, providing valuable insights into the uniformity of our laboratory processes and outcomes.

Individual staff performance analysis

Our analysis of individual embryologists' performance revealed no statistically significant differences ($p > 0.05$) in achieving KPIs (IVF polyspermy rate, ICSI degradation rate, ICSI and IVF fertilization rates, and reasonable blastocyst rate) across selected treatment cycles. Further examination of the distribution of individual embryologist KPIs yielded promising results. The median values for fertilization and blastocyst development rates exceeded the Vienna consensus benchmark for all conducted cycles.[2]

Moreover, these parameters' quarterly values (Q1-Q3) were consistently above the established competency level. This analysis provides a valuable tool for comparing performance across different embryologists within our team. It allows us to identify top performers whose techniques and practices might be shared as best practices. Equally important, it helps us recognize individuals who may benefit from additional support or training, ensuring continuous improvement and maintaining high standards throughout our embryology laboratory.

Time-series analysis

While descriptive statistics provide a quick overview of the central tendencies and spread of KPIs, allowing for easy identification of typical performance levels and outliers, they fall short of capturing the dynamic nature of these metrics over time. We employed moving averages with confidence intervals for each KPI to address this limitation and gain a time-series understanding of changes in KPI shifts. This approach helps identify trends and patterns in KPI performance over time, with the confidence intervals offering insight into the stability and reliability of these trends. We analyzed the laboratory's KPIs stability, focusing on their prognostic influence on pregnancy rates.

Our observations revealed stable oocyte retrieval rates until the third quarter of 2023, followed by a decline in August 2023. Concurrently, we noted a significant decrease ($p < 0.05$) in the blastocyst formation rate, indicating a reduced probability of forming high-quality blastocysts from each retrieved oocyte. The period from early 2022 to the second quarter of 2022 showed a clear trend of increasing oocyte fertilization rates and blastocyst formation rates, correlating with an increased embryo implantation rate observed during this time. A similar trend of increased blastocyst formation rate was observed in 2023. However, while the decrease in blastocyst formation rate in the third quarter of 2023 may appear notable when comparing quarterly averages between years, this decrease is not statistically significant ($p > 0.05$) and amounts to less than 15%.

This finding suggests that no substantial changes in laboratory conditions reduce the overall number of high-quality blastocysts. Our time-series analysis thus provides a more nuanced understanding of KPI fluctuations, allowing us to distinguish between meaningful trends and normal variations in laboratory performance over time.

Regression analysis

We applied ML methods to our dataset to gain a deeper understanding of our data dependencies. After conducting regression analysis using the method of least squares (OLS) for the good blastocyst rate, the following statistical data were obtained: the R-squared (Coefficient of Determination) was 0.393. The F-statistic value was 322.6, and the significance level (F-statistic) was shallow (< 0.0001), leading us to conclude that the model was statistically significant. The coefficient for the number of blastocysts was 0.0670, suggesting the expected change in the frequency of forming blastocysts when the number of blastocysts increases by one. The Durbin-Watson coefficient was close to 2, indicating the absence of significant autocorrelation in the model's residuals and the correctness of the analytical system.

According to the regression analysis, negative dependencies of the blastocyst formation rate on the number of oocytes retrieved and the total number of zygotes were identified, which aligns with analytics from other IVF centers. Notably, the correlations of blastocyst formation rate with the number of retrieved oocytes, the number of inseminated oocytes, and 2pN were not statistically significant ($p = 0.841$, $p = 0.842$, $p = 0.880$, respectively). A substantial ($p < 0.05$) linear dependency was found with the total number of blastocysts. This is an essential indicator of laboratory parameter stability, as an increase in the total number of blastocysts does not deteriorate their quality but increases the probability (proportion) of forming high-quality blastocysts.

KPIs change forecasting

Calculating the laboratory's KPIs during the data preparation procedure allows for the utilization of ML methods to forecast KPI changes with a justified mathematical model. This approach helps determine reference threshold values for quality indicators and identify growth zones for performance improvement. The Seasonal AutoRegressive Integrated Moving Average (SARIMA) model is a time series method widely used for forecasting data with seasonal fluctuations. To achieve this, we integrated the SARIMA model into the laboratory's KPI calculation process to identify the structure of data changes, including seasonal fluctuations, trends, and noise.

This ML model involves determining the orders of autoregression (p), differencing (d), and moving average (q) for the time series components, as well as seasonal orders (P, D, Q). The SARIMA model has been trained on historical data to forecast future performance indicator values based on past observations. These forecasts can calculate target KPIs for the individual laboratory and analyze its effectiveness. Subsequent comparison of forecasted KPIs with actual data forms the basis for a systematic analysis of laboratory efficiency. This approach to internal quality control provides the laboratory with a tool for more predictable process management, early detection of potential issues, and optimization of operations based on data and analytics.

Neural network-based approach

The leading performance indicator of the IVF unit is the clinical pregnancy rate (CPR). The primary challenge in calculating CPR is determining its competence value for each patient population or individual case. According to Maribor's consensus, "Competence and benchmark values of clinical pregnancy rate should be set for a specific local context".[5] Unfortunately, this is not a simple task for an individual IVF center.

Most ML models applied in the field of IVF are based on regression and logistic regression algorithms to identify relationships between the target variable and input parameters.[8] The target variable typically represents the outcome metric of clinic success, such as the frequency of achieving clinical pregnancy, based on fitting data from patient medical history.[9] However, most models used for evaluating CPR are based on patient data and their previous IVF treatment cycles, often failing to track patterns in the changes of quality laboratory parameters relevant to the final transfer outcome.[10]

We developed and implemented our own Deep Neural Network (DNN) within a KPI framework for CPR prediction to address this. This model was trained on our data, with a mean CPR per embryo transfer of 61.93%. The metrics for the model after fitting were: test accuracy = 0.72, AUC = 0.79, PRC = 0.69, precision = 0.72, recall = 0.52, F1 score = 0.61, and MCC = 0.41. After completing the full training process, the predicted CPR was 56.18%, which showed no significant difference ($p = 0.114$) from the actual CPR in our clinic. Utilizing this DNN model, we can compare actual and predictive CPR across time intervals to understand the likelihood of achieving pregnancy. With our neural network model, we established a lower threshold limit for the probability of clinical pregnancy occurrence for each year of operation. A significant difference ($p < 0.05$) was noted for patients in 2021-2022 years compared to 2023 year, with a decrease in the likelihood of clinical pregnancy ranging from 10% at the beginning of the year to 20% after the third quarter of 2023. The theoretically calculated probabilities using the DNN model align with the actual CPR reports in these specified time intervals. In other words, our KPI calculation and DNN model prediction analysis demonstrate that the decrease in CPR from the third quarter of 2023 is a process not directly related to the quality of stimulation (patient preparation) or laboratory work but depends on the patient's initial clinical data.

Using the independent ML algorithm XGBoost, we confirmed the correlation between selected parameters and pregnancy occurrence in our DNN model, highlighting their significance in predicting the outcome of embryo transfer. The parameters identified through linear regression analysis can be considered critical features for developing our DNN for predicting CPR. Utilizing them as the foundation for training the neural network allows for more accurate and reliable forecasting of the probability of clinical pregnancy. This approach ensures a more precise tracing of the mutual influence of laboratory parameters and their impact on the outcome of embryo transfer as part of the internal quality control system. Ultimately, our DNN model will provide clinic staff and patients with more precise and individualized predictions, contributing to more effective infertility treatment outcomes based on laboratory quality and performance indicators.

After training the neural network model, we utilize the probabilities obtained from it and apply logistic regression for calibration. This creates a model that takes outputs from the DNN and calibrates them to probability predictions, improving model performance. As a result, we achieved precision = 0.83 and recall = 0.66.

We chose linear regression to analyze the probabilities predicted by the DNN. The analysis yielded a mean squared error (MSE) value of 0.0027, indicating that our model has a deficient error. This suggests that the model's predictions are close to the actual values of pregnancy probability. The coefficient of determination (R-squared) was found to be 0.937, indicating that approximately 93.7% of the variance in pregnancy probability is explained by the independent variables in the model.

Explanation of DNN results with clustering analysis

To deepen our understanding of the predictive capabilities of the developed DNN for IVF pregnancy probability, we conducted a k-means clustering analysis on the model's predicted output values across a dataset to identify inherent patterns in the expected pregnancy probabilities, offering insights into the diverse patient profiles represented within the dataset. The clustering process partitioned the cases into three distinct clusters, each characterized by unique combinations of predictive features. For the effectiveness analysis of our DNN in predicting pregnancy probability in IVF, we compared predicted clusters with clusters based on accurate data on pregnancy occurrence in our clinic.

This comparison allowed us to identify consistency between predicted and real results and differences that may indicate essential aspects of the data. Both clusters contained similar features characterizing stimulation quality indicators and patient clinical data, such as age, number of follicles, and number of oocytes. We also observed similar trends in the distribution of feature weights characterizing pregnancy probability. This comprehensive analysis demonstrates the robustness of our DNN model in predicting IVF pregnancy probability, providing valuable insights for model explanation and clinical decision-making and enhancing the effectiveness of quality assurance (QA) programs.

DNN validation

To understand the predictive capability of our DNN model, we selected a reference group from our dataset — single embryo transfer of euploid embryos after preimplantation genetic testing for aneuploidy (PGT-A). This selection achieved an initial balance of classes in our patient subpopulation, with an average implantation rate of 54.6%. An analysis of CPR in these protocols revealed a 0.27 error in predicting clinical pregnancy rate, with an AUC = 0.67 (CI = 0.62-0.75) and an accuracy of 0.77. The same metrics (AUC = 0.67) were obtained for embryonic and clinical outcomes of blastocysts with stratified AI scores from iDAScore Embryoscope™ [11] and with other CNN models and commercial AI solutions in IVF: AUC = 0.65 for fresh and frozen transfers and 0.63 for euploid transfers.[12] Consequently, the prediction capability of our DNN model can be used as an entirely noninvasive additional method for embryo selection, similar to other time-lapse systems.[13]

The external validation of our model in 2 independent clinics had comparative results of AUC = 0.73 with a hybrid AI model 3D-ResNet based on several TL technologies (MIRI, GERI, and EMBR-EMBR+) in a cross-validation report from 14 clinics (mean AUC = 0.73) and with video models along (AUC = 0.68) from the same data.[14]

Bayesian method application for prospective approach

However, all these predictions were performed on retrospective data. The Bayesian method offers a robust framework for integrating historical data with real-time predictions from neural networks, enabling a transition from retrospective to prospective analysis. We utilized it to enhance the prediction accuracy of IVF cycle success rates. We defined the prior distribution based on accurate data from IVF cycle successes in 2023-2024. The historical data comprised clinical pregnancy rates and the number of transfers per cycle, transformed into probabilities for each quarter. The quarterly probabilities were as follows: 0.55, 0.55, 0.49, 0.59, and 0.47. A predicted success probability of 0.64 was also obtained from the DNN model for a new cycle.

We initialized our Bayesian model with prior parameters based on historical success and failure counts, specifically with prior successes (α) set to 188 and prior failures (β) set to 106. These priors encapsulate our initial belief about the success rate derived from extensive historical data. Subsequently, we updated our posterior distribution iteratively using the quarterly probabilities. We calculated the equivalent success and failure counts for each probability by scaling the probabilities to a standardized sample size. Following integrating all quarterly probabilities, we incorporated the neural network's predicted success probability for the upcoming cycle. This was done by adjusting the posterior parameters again and adding the expected probability as an additional observation. The final posterior parameters were then used to derive the predicted success rate and associated credible interval for the next IVF cycle. The resulting posterior distribution yielded an expected success rate of approximately 0.58 (58% CPR chance) for the next cycle, with CI = 0.55 - 0.61.

This updated distribution reflects a more accurate estimate of the success rate of the treatment cycle for the future and can help evaluate our observations and CPR expectations. With the dynamic updating of new data, the Bayesian method allows for the seamless integration of historical data and real-time predictions, offering a precise measure of the uncertainty associated with forecasts. This makes the model more robust and adaptive. The combination of the DNN and Bayesian frameworks provides a robust, adaptable, and precise approach to predictive modeling, making it an invaluable tool in the ongoing effort to optimize IVF treatment quality control and data analytics.

Quality management with the help of DNN

The developed neural network model can serve as a unique and comprehensive tool for internal quality control of laboratory parameters and clinic performance by setting lower limits of CPR probability. In many cases, it is challenging to differentiate between patients seeking assistance from various doctors within the clinic, making the specification of competency boundaries in achieving quality targets quite blurry and uncertain. However, our DNN employs a grouped analysis of protocols, allowing their aggregation based on temporal criteria and individual staff members.

We compared the actual clinical pregnancy rates achieved individually by each reproductive specialist during 2023 in external audit programs across four different IVF centers. Through this approach, we identified three doctors (No. 1, No. 2, and No. 3) with actual CPRs of 34.0%, 42.5%, and 40.5%, respectively, which were higher than the theoretically calculated thresholds of 33.60%, 33.33%, and 38.01% for their patient groups ($p > 0.05$). Additionally, we identified three doctors (No. 4, No. 5, and No. 6) whose transfer outcomes (9.60%, 24.82%, 16.01%) require serious monitoring and verification, as their actual CPRs were significantly lower ($p < 0.01$) than those predicted by the model (33.53%, 33.33%, 37.25%) for these staff members. Based on that, it can be concluded that the competency level of doctors No. 4, 5, and 6 still does not allow them to work independently. Every procedure they conduct requires careful monitoring and mentoring from more experienced colleagues, such as specialists No. 1, 2, and 3 within the same center.

Thus, the developed DNN is an integral, accurate, and reproducible element of QA that can be used for both internal QA and external clinic audits, as well as for determining the individual competency of the staff. With its help, it is possible to define the boundaries of personal competency and identify staff members whose work requires increased scrutiny and those who can perform such monitoring. The importance of this analysis lies in the fact that the role of a mentor can be assigned to existing clinic staff without the need for additional external experts or auditors. This significantly simplifies organizational matters and reduces direct financial expenditures on staff training.

CONCLUSION

Our research represents a significant leap forward in applying data analytics to IVF practice. By integrating diverse analytical methods, we have developed a comprehensive framework that goes far beyond traditional descriptive statistics. This approach includes advanced KPI calculations personalized for individual staff members, time series analysis with moving averages, clustering and principal component analysis (PCA) for embryologist performance using machine learning algorithms, linear regression analysis for understanding KPI relationships, and neural network prediction for the result of IVF treatment procedures. This multifaceted approach allows a deeper understanding of laboratory performance, embryologist efficiency, and process stability. It can be applied as a comprehensive quality management system in the embryology laboratory. It serves as a first step in integrating AI in IVF, shifting from the traditional concept of selecting the best embryo to understanding the parameters involved in successful outcomes.

REFERENCES

1. Glatstein I, Chavez-Badiola A, Curchoe CL. New frontiers in embryo selection. *J Assist Reprod Genet.* 2023; 40(2): 223-234. https://doi.org/10.1007/s10815-022-02708-5.
2. ESHRE Special Interest Group of Embryology and Alpha Scientists in Reproductive Medicine. The Vienna consensus: report of an expert meeting on the development of ART laboratory performance indicators. *Reprod Biomed Online.* 2017; 35(5): 494-510. https://doi.org/10.1016/j.rbmo.2017.06.015.
3. Fabozzi G, Cimadomo D, Maggiulli R, Vaiarelli A, Ubaldi FM, Rienzi L. Which key performance indicators are most effective in evaluating and managing an in vitro fertilization laboratory? *Fertil Steril.* 2020; 114(1): 9-15. https://doi.org/10.1016/j.fertnstert.2020.04.054.
4. Bormann CL, Curchoe CL, Thirumalaraju P, et al. Deep learning early warning system for embryo culture conditions and embryologist performance in the ART laboratory. *J Assist Reprod Genet.* 2021; 38(7): 1641-1646. https://doi.org/10.1007/s10815-021-02198-x.
5. ESHRE Clinic PI Working Group, Vlaisavljevic V, Apter S, et al. The Maribor consensus: report of an expert meeting on the development of performance indicators for clinical practice in ART. *Hum Reprod Open.* 2021; 2021(3): hoab022. https://doi.org/10.1093/hropen/hoab.
6. Zacà C, Coticchio G, Vigiliano V, et al. Fine-tuning IVF laboratory key performance indicators of the Vienna consensus according to female age. *J Assist Reprod Genet.* 2022; 39(4): 945-952. https://doi.org/10.1007/s10815-022-02468-2.
7. Khan HL, Boothroyd C, Chang TA, et al. ASPIRE Guidelines for Assisted Reproductive Technology (ART) Laboratory Practice in Low and Medium Resource Settings. *Fertil Reprod.* 2023; 5(3): 115-133.
8. Alizadehsani R, Roshanzamir M, Hussain S, et al. Handling of uncertainty in medical data using machine learning and probability theory techniques: a review of 30 years (1991-2020). *Ann Oper Res.* 2021; 1-42. https://doi.org/10.1007/s10479-021-04006-2.
9. Hernández-González J, Inza I, Crisol-Ortíz L, et al. Fitting the data from embryo implantation prediction: Learning from label proportions. *Stat Methods Med Res.* 2018; 27(4): 1056-1066. https://doi.org/10.1177/0962280216651098.
10. Li L, Cui X, Yang J, Wu X, Zhao G. Using feature optimization and LightGBM algorithm to predict the clinical pregnancy outcomes after in vitro fertilization. *Front Endocrinol (Lausanne)*. 2023; 14: 1305473. https://doi.org/10.3389/fendo.2023.1305473.
11. Bamford T, et al. A comparison of morphokinetic models and morphological selection for prioritizing euploid embryos: a multicentre cohort study. *Hum Reprod.* 2024; 39: 53-61. https://doi.org/10.1093/humrep/dead237.
12. Loewke K, Cho JH, Brumar CD, et al. Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos. *Fertil Steril.* 2022; 117(3): 528-535. https://doi.org/10.1016/j.fertnstert.2021.11.022.

13. Enatsu N, Miyatsuka I, An LM, et al. A novel system based on artificial intelligence for predicting blastocyst viability and visualizing the explanation. *Reprod Med Biol.* 2022; 21(1): e12443. https://doi.org/10.1002/rmb2.12443.
14. Duval A, Nogueira D, Dissler N, et al. A hybrid artificial intelligence model leverages multi-centric clinical data to improve fetal heart rate pregnancy prediction across time-lapse systems. *Hum Reprod.* 2023; 38(4): 596-608. https://doi.org/10.1093/humrep/dead023.

Conflict of Interest: The authors declare no conflicts of interest relevant to this study.