RBMO

COMMENTARY

Empathetic application of machine learning may address appropriate utilization of ART

Julian Jenkins[1,2,*], Sheryl van der Poel[3], Jan Krüssel[4], Ernesto Bosch[5], Scott M. Nelson[6], Anja Pinborg[7], Mylene M.W. Yao[8]

ABSTRACT

The value of artificial intelligence to benefit infertile patients is a subject of debate. This paper presents the experience of one aspect of artificial intelligence, machine learning, coupled with patient empathy to improve utilization of assisted reproductive technology (ART), which is an important aspect of care that is under-recognized. Although ART provides very effective options for infertile patients to build families, patients often discontinue ART when further treatment is likely to be beneficial and most of these patients do not achieve pregnancy without medical aid. Use of ART is only in part dependent on financial considerations; stress and other factors play a major role, as shown by high discontinuation rates despite reimbursement. This commentary discusses challenges and strategies to providing personalized ART prognostics based on machine learning, and presents a case study where appropriate use of such prognostics in ART centres is associated with a trend towards increased ART utilization.

INTRODUCTION

The potential of assisted reproductive technology (ART) to address infertility is limited by the under-utilization of well-established ART, where appropriate and legally allowed. This ART under-utilization is inadequately captured at national level and its impact on cumulative pregnancy rates is under-appreciated. Although financial considerations are an obvious barrier to the use of ART, even with full reimbursement, patients appear reluctant to move onto and persist with ART for as long as would be anticipated to be beneficial. This suggests stress and other factors are key barriers to ART utilization. One patient-centred approach to addressing stress is to clearly communicate individualized prognostic information to patients with transparency and empathy, thereby helping patients set realistic expectations of ART. Shared decision-making may also reduce the stress of healthcare practitioners counselling patients on ART prognosis.

This commentary discusses challenges to providing accurate, personalized ART prognostics with any modelling method, challenges and potential benefits specific to the use of machine learning. Finally, we present a case study (Univfy), sharing strategies to address challenges of ART prognostics and ongoing efforts to improve empathetic prognostics counselling benefiting appropriate ART under-utilization.

LIMITED ART UTILIZATION COMPROMISES PER-PATIENT SUCCESS

Premature ART discontinuation commonly occurs even in countries providing national reimbursement, negatively impacting ultimate per-patient live birth rates. An analysis of 122,560 couples undergoing ART in Germany revealed that 37,220 (30.4%)

KEYWORDS
ART utilization
Artificial intelligence
IVF drop-outs
Machine learning
Patient empathy
Prognostication

[1] Repromed Sàrl, Crassier 1263, Switzerland
[2] Medical Affairs, Gedeon Richter Plc / PregLem SA, 41A Route de Frontenex, Geneva 1207, Switzerland
[3] Population Council, One Dag Hammarskjold Plaza, New York, NY 10017, USA
[4] Department of Obstetrics and Gynecology, Heinrich-Heine University Medical Centre, Düsseldorf, Germany
[5] Instituto Valenciano de Infertilidad (IVI-RMA), Valencia, Spain
[6] School of Medicine, University of Glasgow, Glasgow G31 2ER, UK
[7] Fertility Clinic, Section 4071, Rigshospitalet, Copenhagen University Hospital, Blegdamsvej 9, DK-2100, Copenhagen, Denmark
[8] Univfy Inc., 1 1st Street, Los Altos, CA, USA

*Corresponding author. E-mail address: julian.jenkins@repromed.net (J Jenkins). https://doi.org/10.1016/j.rbmo.2020.07.005 1472-6483/
Declaration: Mylene Yao is the Chief Executive Officer of Univfy Inc. and has a Univfy IP Portfolio US Patent Number 9,458,495B2, foreign counterparts and other issued and pending patents.

discontinued after the first or second unsuccessful treatment (*Kreuzer et al., 2018*). The Danish national registers revealed within 5 years of starting treatments with ART of 2137 women, 53% delivered after ART, 11% delivered after achieving pregnancy without medical aid and 0.6% delivered after intrauterine insemination (*Malchau et al., 2017*). As Denmark has one of the highest ART utilization rates in Europe this may represent the best-case scenario (*De Geyter et al., 2018*).

LIMITATIONS OF ART REGISTRY DATA FOR COUNSELLING PATIENTS

National registries have limitations for ART prognostics counselling, because this is not their aim. Prior to consulting specialists, patients could learn about ART success rates from online calculators such as those developed from UK or US registries (https://w3.abdn.ac.uk/clsm/opis or https://www.sartcorsonline.com/Predictor/Patient). However, patients consulting fertility specialists are better counselled by informed centre-specific ART prognostics. Registries do not collect all variables important for prognostication, clinical practice differs between ART clinics and even terminology may be used inconsistently (*Jenkins et al., 2004*). Prognostic models need sufficient relevant data specificity for both patients and the chosen treatment centre.

CHALLENGES IN PREDICTION OF PREGNANCY FOLLOWING ART

Many ART outcome prediction models have been reported, although they are not widely adopted and their impact is unknown (*Ratna et al., 2020*). The relevant literature has primarily focused on technical issues rather than clinical utility. In prognostic modelling, broadly speaking, there are two main types of errors – bias or 'underfitting' and variance or 'overfitting'. Imagine first a prediction model using a data set comprising 100,000 ART cycles and only one clinical variable, age. This model would have significant underfitting, i.e. failing to 'see' relationships between many potentially relevant variables and outcomes. Next, imagine a prediction model using a data set comprising 100 ART cycles, and 50 clinical variables. This prediction model is likely to have significant overfitting: over-interpreting noise as relationships between variables and outcomes. The optimal prediction model minimizes both errors, settling on the appropriate 'bias and variance trade-off' to best serve the clinical goals.

Although the patient's age is often the top prognostic factor, age accounts for only ~50% of the prediction prior to starting an IVF cycle. Whereas it is simple for fertility clinics to present success rates against age, it becomes increasingly difficult to include additional factors to improve prognostication. The development of sound centre-specific prediction models requires expertise typically unavailable at ART centres. Common challenges of ART prognostic modelling include technical or design flaws in the original modelling work (*Christodoulou et al., 2019*; *Leushuis et al., 2009*). Further, although models may be used beyond their originating population, validation in the new population is required and may fail due to diverse patient clinical profiles, non-uniform use of diagnostic testing and criteria, and variations in treatment ovarian stimulation protocols and embryology protocols at other centres (*Christodoulou et al., 2019*; *Leushuis et al., 2009*). Separately, there is a paucity of discussion in the literature about the requirements of real-world usage of ART prognostics counselling beyond academic research. Prediction models requiring variables available only after starting ART render them less relevant to patients considering their first ART cycle. Even if a reliable centre-specific model were developed, its quality-assured implementation into an easily usable tool presents additional challenges.

CONVERGENCE OF DEVELOPMENTS ENABLING MACHINE LEARNING PROGNOSTICATION IN ART

One approach to address the above challenges is through machine learning, in view of technological advances and changes in willingness to trust prediction models. For many years, web-based knowledge sharing has supported both public and professional education (*Jenkins et al., 1999*; *Whittington et al., 2004*) and now machine learning is increasingly being applied to ART (*Zaninovic et al., 2019*). Cloud computing, internet speed and ubiquity, versatility of programming frameworks and security technology have made it feasible to provide personalized ART machine learning prognostication services, although machine learning in itself is not a 'magic bullet', requiring expert application, strict validation and meaningful presentation of probabilities to patients.

APPLYING MACHINE LEARNING TO ART PROGNOSTICATION

Machine learning is an application of artificial intelligence often described as a branch of computing focusing on models that automatically learn from structured data, such as held in ART databases. Logistic regression and linear regression are earlier examples of supervised machine learning algorithms, whereas 'traditional machine learning' or just 'machine learning' often refers to more advanced machine learning methods such as Bayesian network, support vector machine, gradient boosting machine (GBM) or random forest. Deep learning artificial intelligence applications are part of broader machine learning most useful when analysing very large datasets, especially of unstructured information, such as images of blastocysts to select for transfer, and is not the focus of the discussion here. Thus, the broader machine learning heading comprises methodologies (logistic regression, traditional machine learning, and deep learning) that can address a wide range of data sets. There is no need to advocate for one method or another, and modelling techniques may even be combined. For example, when modelling the ART dataset described in *Nelson et al. (2015)*, the prediction model resulting from combined GBM and logistic regression in the form of generalized linear method (GLM) was superior to models developed from using GBM alone or GLM alone. However, that does not mean that each ART dataset should be treated this way and objective assessments are needed to determine the most appropriate technique for achieving the best prediction model for the clinical goals. One potential benefit of machine learning, such as GBM, in ART prognostication is that it provides a framework for using regression in a way that allows the comprehensive testing of training and test set designs and variable selection to be performed with efficiency and this framework, once established, can be applied to many distinct data sets originating from different sources (*Friedman, 2001*). The caveat is that combined machine learning modelling

expertise and an in-depth understanding of potential biases in clinical ART data are required to ensure proper design and feature engineering (the selection of variables and their formats) to avoid overfitting and unintended biases (*Vollmer et al., 2018*).

CASE STUDY: A REAL-WORLD APPLICATION OF ARTIFICIAL INTELLIGENCE TO SUPPORT PATIENT COUNSELLING IN ART

By using a platform that achieved scalability from having analysed over 200,000 ART cycles from over 50 ART centres across North America, Europe and Asia, Univfy enables ART success prognostic models to be trained and validated on a centre-specific basis (Univfy® AI Platform, issued and pending US and global patents). The platform has addressed many clinical needs, for which performance metrics have been published, including the use of first ART cycle data to predict a subsequent ART cycle live birth probability (*Banerjee et al., 2010*), the use of clinical data prior to the first ART cycle to predict first ART cycle live birth probability using both a multicentre (*Choi et al., 2013*) and a centre-specific approach (*Nelson et al., 2015*), and the use of clinical data available on the day of embryo transfer to predict the probability of multiple births from double embryo transfer for elective single-embryo transfer counselling (*Lannon et al., 2012*).

Now new centres may be taken through a series of steps to develop, validate and deploy predictive models to support their clinical care. The first step is to meet with the centre's leadership to define patient counselling goals. In compliance with relevant privacy and healthcare laws and other confidentiality conditions of the collaboration agreement, the centre submits relevant encrypted data to secure servers for processing into a proprietary, analysis-ready format for modelling. The data set typically comprises 1 to 5 years of ART cycles with linked outcomes from fresh and frozen embryo transfers, relevant clinical variables such as age, body mass index, ovarian reserve, reproductive history, clinical diagnosis and male partner's health data, semen analysis and clinical outcomes (e.g. clinical pregnancy, live birth).

The data science team analyses the data to identify any special features that together with clinicians' feedback and operational requirements will determine strategies of features engineering, training and test set design, modelling techniques and the use of an optimal number or combinations of variables tested to avoid over- or underfitting given the size of the training set. Following partitioning of the data into training and test sets, machine learning such as GBM is applied to generate hundreds of models from the training sets and model performance metrics are obtained by validating against test sets. After comparing against the age-based control, models are compared and the model that maximizes model performance and reproducibility and best meets clinical utility considerations is ultimately selected. For example, prognostics models developed for smaller data sets (with ~250 ART cycles on the lower end) would typically still outperform corresponding age-based control models, although the model would use fewer clinical variables to avoid overfitting and the extent of personalization would be less than models developed using larger data sets. As outcomes of more ART cycles become available, the model can be updated to utilize more clinical variables.

Model performance is assessed by predictive power (posterior log of odds ratio compared with age, PLORA), discrimination (AUC), and reclassification compared against age-only prognostics models according to published methods (*Banerjee et al., 2010*; *Nelson et al., 2015*). **TABLE 1** illustrates the benefits of machine learning predictive model reclassification versus age alone. Prognostics model development and validation are followed by comprehensive quality assurance testing, establishment of workflow and patient flow incorporating use of the personalized prognostics report, and role-specific staff training prior to clinical usage.

The Univfy PreIVF Report for patients conveys known prediction errors and uncertainties not addressable by the prognostics model so that patients are aware and physicians can efficiently and easily explain them. It is important from the outset in counselling patients to stress prognostication limitations, as clearly it is impossible to know the absolute outcome or the cause or circumstances of ART cycle failure (e.g. no embryo to transfer versus no implantation) for a particular patient in advance. In most cases the embryology results of the first treatment cycle are consistent with predicted probability of ART success, hence has little additional impact on the probability of success in the subsequent cycles over other clinical factors, allowing prognostication for three or more cycles. However, in

TABLE 1 COMPARISON OF TIERS OF PREDICTED PROBABILITY OF LIVE BIRTH ACCORDING TO AGE COHORT GROUPS VERSUS RECLASSIFICATION INTO GROUPS ACCORDING TO A BOOSTED TREE METHOD TO TRAIN A PREIVF DIVERSITY PREDICTIVE MODEL USING 1061 FIRST IVF CYCLES VALIDATED WITH A FURTHER 1058 FIRST IVF CYCLES

Tiers of predicted probability of live birth	Age cohort groups	Proportion of patients in each age group	PreIVF D classification groups	Proportion of patients in each group	Impact of PreIVF D reclassification
>45.0%	n/a		Group A	41.6%	• Around 86% of patients had significantly different probabilities of live birth than predicted by age alone ($P < 0.05$) • Around 57% of patients had higher and 28% patients had lower probabilities of live birth • Predictive power by log-likelihood increased by 36% (or increased by 9.0-fold on the log scale)
40.0–45.0%	<35 years	49.6%	Group B	13.5%	
30.0–39.9%	35–37 years	24.2%	Group C	19.3%	
20.0–29.9%	38–40 years	19.1%	Group D	13.2%	
10.0–19.9%	n/a		Group E	7.3%	
<10.0%	41–42 years	7.1%	Group F	5.1%	

Adapted from *Choi et al. (2013)*.

n/a = not appropriate; PreIVF D = PreIVF diversity.

2–5% of cases, the use of oocyte or embryo morphology data from the failed cycle informs a much altered probability of success in subsequent cycles and patients need to be aware of this possibility. Also, some patients may have unique circumstances impacting ART prognosis yet they are not well represented in the centre's historical ART outcomes data set, hence physicians need to assess the suitability of patients prior to applying predictive models.

Many patients may not expect that multiple ART cycles might be required for a reasonable probability of having a baby, assuming that expensive, 'advanced' treatment will achieve a baby on the first attempt. Such misconceptions cause unrealistic expectations, greater psychological burden and increase discontinuation rates after one failed treatment, even if the patient has good prognosis.

With the support of the Univfy PreIVF Report, personalized to the couple's health data and developed and validated against the fertility centre's outcomes data, physicians continue to apply their professional knowledge when counselling their patients, with enhanced effectiveness backed by visual and data-centric formats. For this commentary a utilization analysis of anonymized, aggregated Univfy PreIVF Report data representing ~61,000 new patient visits and their subsequent treatment choices indicated that from the introduction of counselling supported by this machine learning-driven report, there was a nearly two-fold increase (1.8-fold increase, 95% confidence interval 1.2-fold to 3.5-fold) in progression to ART within 6 months from new patient visits compared with historical data at each fertility treatment centre.

CONCLUSION

Today ART provides options for almost all infertile patients to have families, but to realize the potential, patients need to commit to a course of treatments to maximize per-patient success. Where financial barriers exist, accurate prognostication across multiple treatment allows IVF centres to offer personally tailored IVF refund programmes or multicycle programmes to the majority of patients, reducing financial uncertainty and conveying the commitment and confidence of IVF centres. Where there are no financial barriers, accurate prognostication is a critical tool for counselling patients as to whether ART may be appropriate and to set realistic expectations reducing the emotional rollercoaster of treatment, ultimately supporting more patients to establish families. When executed appropriately, the use of artificial intelligence/machine learning has the potential to improve the personalization and empathy of ART prognostics counselling for physicians and patients. A collaborative approach taken by expert modellers and ART practitioners can strike a healthy balance between critiquing and trusting artificial intelligence to maximize the potential of this technology to improve appropriate ART utilization.

ACKNOWLEDGEMENTS

We thank A Xie, S Lin and R Thatcher for the Univfy® PreIVF Report utilization analysis presented in the paper and Professor WH Wong for reviewing the manuscript.

REFERENCES

Banerjee, P., Choi, B., Shahine, L.K., Jun, S.H., O'Leary, K., Lathi, R., Westphal, L.M., Wong, W.H., Yao, M.W.M. **Deep phenotyping to predict live birth outcomes in *in vitro* fertilization.** Proc. Natl. Acad. Sci. USA 2010; 107: 13570–13575

Choi, B., Bosch, E., Lannon, B.M., Leveille, M.C., Wong, W.H., Leader, A., Pellicer, A., Penzias, A.S., Yao, M.W.M. **Personalized prediction of first-cycle *in vitro* fertilization success.** Fertil. Steril. 2013; 99: 1095–1911

Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y., Van Calster, B. **A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models.** J. Clin. Epidemiol. 2019; 110: 12–22

De Geyter, C., Calhaz-Jorge, C., Kupka, M.S., Wyns, C., Mocanu, E., Motrenko, T., Scaravelli, G., Smeenk, J., Vidakovic, S., Goossens, V. **European IVF-monitoring Consortium (EIM) for the European Society of Human Reproduction and Embryology (ESHRE). ART in Europe, 2014: results generated from European registries by ESHRE: The European IVF-monitoring Consortium (EIM) for the European Society of Human Reproduction and Embryology (ESHRE).** Hum. Reprod. 2018; 33: 1586–1601

Friedman, J.H. **Greedy function approximation: a gradient boosting machine.** Ann. Stat. 2001; 29: 1132–1189

Jenkins, J.M. **The Internet, intranets and Reproductive Medicine.** Hum. Reprod. 1999; 14: 586–589

Jenkins, J., Daya, S., Kremer, J., Balasch, J., Barratt, C., Cooke, I., Lawford-Davies, J., De Sutter, P., Suikari, A.M., Neulen, J., Nygren, K. **European Classification of Infertility Taskforce (ECIT) response to Habbema. Towards less confusing terminology in reproductive medicine: a proposal.** Hum. Reprod. 2004; 19: 2687–2688

Kreuzer, V.K., Kimmel, M., Schiffner, J., Czeromin, U., Tandler-Schneider, A., Krüssel, J.S. **Possible Reasons for Discontinuation of Therapy: an Analysis of 571 071 Treatment Cycles From the German IVF Registry.** Geburtsh Frauenheilk 2018; 78: 984–990

Lannon, B.M., Choi, B., Hacker, M.R., Dodge, L.E., Malizia, B.A., Barrett, C.B., Wong, W.H., Yao, M.W.M., Penzias, A.S. **Predicting personalized multiple birth risks after *in vitro* fertilization double embryo transfer.** Fertil. Steril. 2012; 98: 69–76

Leushuis, E., van der Steeg, J.W., Steures, P., Bossuyt, P.M., Eijkemans, M.J., van der Veen, F., Mol, B.W., Hompes, P.G. **Prediction models in reproductive medicine: a critical appraisal.** Hum. Reprod. 2009; 15: 537–552

Malchau, S.S., Henningsen, A.A., Loft, A., Rasmussen, S., Forman, J., Nyboe Andersen, A., Pinborg, A. **The long-term prognosis for live birth in couples initiating fertility treatments.** Fertil. Steril. 2017; 32: 1439–1449

Nelson, S.M., Fleming, R., Gaudoin, M., Choi, B., Santo-Domingo, K., Yao, M. **Antimullerian hormone levels and antral follicle count as prognostic indicators in a personalized prediction model of live birth.** Fertil. Steril. 2015; 104: 325–332

Ratna M, B., Bhattacharya, S., Abdulrahim, B., McLernon D, J. **A systematic review of the**

quality of clinical prediction models in *in vitro* fertilization. Hum. Reprod. 2020; 35: 100–116

Vollmer S., Mateen B.A., Bohner G., Kiraly F.J., Ghani R., Jonsson P., Cumbers S., Jonas A., McAllister K.S.L., et al. Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. arXiv:1812.1040v1 2018. https://arxiv.org/abs/1812.10404

Whittington, K., Cook, J., Barratt, C., Jenkins, J. **Can the Internet widen participation in reproductive medicine education for professionals?** Hum. Reprod. 2004; 19: 1800–1805

Zaninovic, N.O., Elemento, O., Rosenwaks, Z. **Artificial intelligence: its applications in reproductive medicine and the assisted reproductive technologies.** Fertil. Steril. 2019; 112: 28–30

Received 16 February 2020; received in revised form 27 April 2020; accepted 9 July 2020.