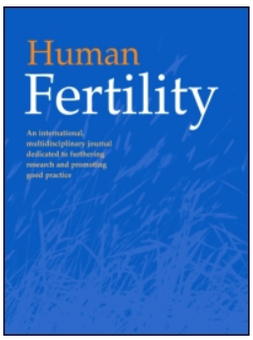

Taylor & Francis
Taylor & Francis Group

Human Fertility

an international, multidisciplinary journal dedicated to furthering research and promoting good practice

ISSN: 1464-7273 (Print) 1742-8149 (Online) Journal homepage: http://www.tandfonline.com/loi/ihuf20

Assessing efficacy of day 3 embryo time-lapse algorithms retrospectively: impacts of dataset type and confounding factors

Yanhe Liu, Katie Feenan, Vincent Chapple & Phillip Matson

To cite this article: Yanhe Liu, Katie Feenan, Vincent Chapple & Phillip Matson (2018): Assessing efficacy of day 3 embryo time-lapse algorithms retrospectively: impacts of dataset type and confounding factors, Human Fertility, DOI: 10.1080/14647273.2018.1425919

To link to this article: https://doi.org/10.1080/14647273.2018.1425919

Published online: 16 Jan 2018.

HUMAN FERTILITY, 2018
https://doi.org/10.1080/14647273.2018.1425919

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Assessing efficacy of day 3 embryo time-lapse algorithms retrospectively: impacts of dataset type and confounding factors

Yanhe Liu[a,b], Katie Feenan[a], Vincent Chapple[a] and Phillip Matson[a,b]

[a]Fertility North, Joondalup, Australia; [b]School of Medical and Health Sciences, Edith Cowan University, Joondalup, Australia

ABSTRACT

This study investigated the efficacy of four published day 3 embryo time-lapse algorithms based on different types of datasets (known implantation data [KID] and single embryo transfer [SET]), and the confounding effect of female age and conventional embryo morphology. Four algorithms were retrospectively applied to three types of datasets generated at Fertility North between February 2013 and December 2014: (a) KID dataset ($n = 270$), (b) a subset of SET ($n = 144$, end-point = implantation), and (c) SET ($n = 144$, end-point = live birth), respectively. All four algorithms showed progressively reduced predictive power (expressed as area under the receiver operating characteristics curve and 95% confidence interval [CI]) after application to the three datasets (a–c): Liu (0.762 [0.701–0.824] vs. 0.724 [0.641–0.807] vs. 0.707 [0.620–0.793]), KIDScore (0.614 [0.539–0.688] vs. 0.548 [0.451–0.645] vs. 0.536 [0.434–0.637]), Meseguer (0.585 [0.508–0.663] vs. 0.56 [0.462–0.658] vs. 0.549 [0.445–0.652]), and Basile (0.582 [0.505–0.659] vs. 0.519 [0.421–0.618] vs. 0.509 [0.406–0.612]). Furthermore, using KID dataset, the association (expressed as odds ratio and 95% CI) between time-lapse algorithms and implantation outcomes lost statistical significance after adjusting for conventional embryo morphology and female age in 3 of the 4 algorithms (KIDScore 1.832 [1.118–3.004] vs. 1.063 [0.659–1.715], Meseguer 1.150 [1.021–1.295] vs. 1.122 [0.981–1.284] and Basile 1.122 [1.008–1.249] vs. 1.038 [0.919–1.172]). In conclusion, SET is a preferred dataset to KID when developing or validating time-lapse algorithms, and day 3 conventional embryo morphology and female age should be considered as confounding factors.

ARTICLE HISTORY
Received 15 August 2017
Accepted 11 December 2017

KEYWORDS
Implantation; single embryo transfer (SET); embryo viability

Introduction

Embryo selection is considered as one of the most crucial components in *in vitro* fertilization (IVF) treatment, given a large proportion of human embryos created *in vitro* fail to implant following uterine transfer (Cohen, Alikani, & Bisignano, 2012). A number of embryo selection approaches, both invasive and non-invasive, have been developed in the last three decades with some clinical applications (Montag, Toth, & Strowitzki, 2013). However, their effectiveness is yet to be fully validated for identifying embryos with the highest implantation potential from a given cohort (Gardner & Balaban, 2016; Gardner, Meseguer, Rubio, & Treff, 2015). Blastocyst transfer is widely applied in IVF clinics for further selection of cleavage stage embryos, by extending culture *in vitro* (Gardner et al., 1998). Nonetheless, concerns over the potential epigenetic impact on the subsequent offspring following extended culture have been reported by a number of groups (Dar, Lazer, Shah, & Librach, 2014; Kallen et al., 2010; Kalra, Ratcliffe, Barnhart, & Coutifaris, 2012; Maheshwari, Kalampokas, Davidson, & Bhattacharya, 2013), but this debate is still ongoing (Gardner, 2016; Maheshwari, Hamilton, & Bhattacharya, 2016; Martins et al., 2016).

The recent introduction of time-lapse technology appeared to be a promising tool in optimizing embryo selection at the cleavage stage, due to its non-invasive nature, uninterrupted culture environment, and the large amounts of dynamic morphological information captured (Meseguer et al., 2011; Wong et al., 2010). Several embryo selection algorithms showed encouraging results (Basile et al., 2015; Conaghan et al., 2013; Liu, Chapple, Feenan, Roberts, & Matson, 2016; Meseguer et al., 2011; Milewski et al., 2015; Petersen, Boel, Montag, & Gardner, 2016); although well-designed, high quality randomized controlled trials (RCTs) are still lacking (Kaser & Racowsky, 2014; Kirkegaard et al., 2014; Racowsky, Kovacs, & Martins, 2015). The difficulty of managing such RCTs is widely acknowledged in terms of their design and initiation,

CONTACT Yanhe Liu gift0409@yahoo.com.au Reproductive Medicine Center, Tianjin United Family Hospital, 22 Tianxiaoyuan, Tanjiang Road, Hexi District, Tianjin 300221, China

as such evidence generated via retrospective analysis still remains important (providing confounding factors are adequately addressed through statistical analysis) (Kirkegaard et al., 2016). Currently, the vast majority of reported time-lapse algorithms are based on retrospective blastulation results or known implantation data (KID) with limited subsequent RCT validation. Furthermore, potentially confounding factors, such as female age and conventional embryo morphology, have possibly gone unaddressed in the statistical analysis.

This study aims to investigate the predictive power of day 3 embryo time-lapse algorithms: (a) when using KID dataset compared to a subset of single embryo transfer (SET) cycles using implantation or live birth as the end-point; and (b) when the confounding effects of female age and conventional embryo morphology are taken into consideration.

Materials and methods

Patient management

This study included a total of 212 IVF and intracytoplasmic sperm injection (ICSI) treatment cycles performed at Fertility North between February 2013 and December 2014. A total of 270 resulting KID embryos, as previously defined (Liu, Chapple, Roberts, & Matson, 2014), were included for analysis. A subset of 144 SET cycles was further analysed following the removal of repeat cycles by the same patients (Table 1). All patients consented to the use of the Embryoscope™ (Vitrolife, Goteborg, Sweden) as the embryo incubation device. Retrospective analysis of data was approved by the Ethics Committees at both Joondalup Health Campus and Edith Cowan University. Clinical procedures including ovarian stimulation, gamete collection and insemination were performed as per previously described (Liu et al., 2014).

Embryo culture and annotation

All normally fertilized oocytes (defined by visualization of 2 pronuclei) were cultured in the Embryoscope™ under 37 °C with 6% CO_2, 5% O_2 in N_2 until day 3 before transfer, cryopreservation, or being allowed to perish. Time-lapse annotation of embryos, including both quantitative and qualitative parameters, was performed by one embryologist using the EmbryoViewer® software. Start time point (t0) was defined as either the time of mixing sperm and oocytes in conventional IVF cases or the middle time point of the sperm injection process for entire oocyte cohort in ICSI cases (Ciray et al., 2014). Quantitative measures included time from insemination to pronuclear fading (tpnf), 2- (t2), 3- (t3), 4- (t4), 5- (t5), 6- (t6), 7- (t7) and 8-cell (t8) stages (Ciray et al., 2014). Relative timings were also calculated, including t5_pnf (t5-tpnf), t3_pnf (t3-tpnf), (t5–t3)/(t5–t2), cc2 (t3–t2) and s2 (t4–t3) (Liu et al., 2016; Liu, Chapple, Feenan, Roberts, & Matson, 2015b; Petersen et al., 2016). Qualitative parameters included direct cleavage (DC), reverse cleavage (RC), evenness at the 2-cell stage, multinucleation and intercellular contact points at the 4-cell stage (Basile et al., 2015; Liu et al., 2014; Liu et al., 2016; Liu, Chapple, Feenan, Roberts, & Matson, 2015a; Meseguer et al., 2011; Rubio et al., 2012).

Table 1. Summary of datasets analysed.

	Dataset	
Parameters	KID	SET (subset)
No. of cycles	212	144
Female age (years, mean ± SD)	34.5 ± 4.5	33.9 ± 4.6
Proportion of IVF/ICSI (%)	39.6/60.4	41.7/58.3
No. clinical pregnancies (%)	61 (28.8)	50 (34.7)
No. live births (%)	45 (21.2)	42 (29.3)

KID: known implantation data; SET: single embryos transfer.

Embryo selection and transfer and treatment outcomes

Either one or two embryos, dependent upon female age and prognosis judged by the treating medical practitioner, with the best conventional morphology from the same cohort were selected for transfer after 3 d culture (Alpha Scientists in Reproductive Medicine & ESHRE Special Interest Group of Embryology, 2011). Briefly, embryos are graded at 68 ± 1 h post insemination, using cell number and fragmentation/evenness (e.g. 8C1), where fragmentation and evenness are considered together ranging from 1 to 4. Top quality embryos are those ≥8C1, good quality embryos ≥8C2 or 6–7C1, fair quality embryos ≥8C3 or 6–7C2 or 5C1, and the remaining are deemed poor quality embryos. Selection was primarily based on day 3 morphology, then day 2 if similar day 3 grades, then multinucleation at the 2- or 4-cell stage if selection still cannot be made. Implantation was confirmed by the detection of foetal heartbeat under ultrasound at 7 weeks of pregnancy and all pregnancies were followed up until birth.

Retrospective embryo grading using different algorithms

Four published time-lapse algorithms (Basile et al., 2015; Liu et al., 2016; Meseguer et al., 2011; Petersen et al., 2016), which had been developed using KID analysis of day 3 embryos cultured in the

Table 2. Comparison of features between 4 time-lapse day 3 embryo selection algorithms based on known implantation outcomes following embryo culture in the Embryoscope™.

	Day 3 time-lapse embryo selection algorithms			
	Meseguer et al. (2011)	Basile et al. (2015)	KIDScore (Petersen et al., 2016)	Liu et al. (2016)
Qualitative time-lapse measures	MN at 4C, evenness at 2C, DC at 1C	MN at 4C, evenness at 2C, DC at 1C	N/A but may be partially addressed by using relative timing expressions	RC (at 1–4C), DC (at 1–4C), <6 ICCP at the end of 4C
Quantitative measures	t5, s2, cc2	t3, t5, cc2	(t3–tpnf), t3, (t5–t3)/(t5–t2)	t5_pnf, s2
Use of D3 conventional morphological score	Discard criteria	Discard criteria	cell count at 66 h	Discard criteria, and cell count at 68 h
Insemination method	ICSI	ICSI	IVF/ICSI ($t0$ = insemination)	IVF/ICSI ($t0$ = pnf)
Resulting grades	A+, A−, B+, B−, C+, C−, D+, D−, E, F	A+, A−, B+, B−, C+, C−, D+, D−, E, F	1, 2, 3, 4, 5	A+, A, B, C, D, E, F
Model development	Single centre retrospective	Multi-centre retrospective	Multi-centre retrospective	Single centre retrospective
Predictive power reported AUC (95% CI)	0.720 (0.645–0.795)	0.610 (0.574–0.638)	0.650 (n/a)	0.762 (0.701–0.824)
Internal validation	RCT, implantation as end-point (Rubio et al., 2014).	Retrospective, multi-centre, implantation as end-point (Basile et al., 2015).	Retrospective, multi-centre, blastulation and blastocyst quality as end-point (Petersen et al., 2016).	Prospective, implantation as end-point (Liu et al., 2016).
External validation	Retrospective, single or multiple centre; blastulation, or blastocyst quality, or implantation as end-point (Best et al., 2013; Freour et al., 2015; Liu et al., 2015; Petersen et al., 2016; Yalcinkaya et al., 2014).	Retrospective, single centre, implantation as end-point (Barrie et al., 2017); Retrospective, multi-centre, blastulation and blastocyst quality as end-points (Petersen et al., 2016).	N/A	Retrospective, multi-centre, blastulation and blastocyst quality as end-points (Petersen et al., 2016).

MN: multinucleation; DC: direct cleavage; RC: reverse cleavage; ICCP: intercellular contact points; t5: duration between insemination and 5C stage; s2: duration of 3C stage; cc2: duration of 2C stage; t3: duration between insemination and 3C stage; tpnf: duration between insemination and pronuclear fading; t2: duration between insemination and 2C stage; t5_pnf: duration between pronuclear fading and 5C stage; AUC: area under the curve; CI: confidence interval; RCT: randomized controlled trial.

Embryoscope™, were retrospectively applied to the dataset (the dataset that was originally used to develop the Liu algorithm (Liu et al., 2016)). The dataset included 270 KID embryos that had developed to at least the 5-cell stage on day 3. After applying the four algorithms, embryos were classified into 10 categories (A+, A−, B+, B−, C+, C−, D+, D−, E and F) according to Meseguer et al. (2011) or Basile et al. (2015); five categories (1, 2, 3, 4 and 5) according to KIDScore (Petersen et al., 2016); and seven categories (A+, A, B, C, D, E and F) according to Liu et al. (2016). Detailed comparisons between the four algorithms are shown in Table 2.

Statistical analysis

Predictive power of the four algorithms was assessed by the receiver operating characteristics (ROC) analysis expressed as area under the ROC curve (AUC, 95% confidence interval [CI]) on (a) the full dataset of 270 KID embryos using implantation as the end-point, and (b) a subset of SETs with repeat patient cycles removed and using either implantation or live birth as the end-point. Additionally, multiple logistic regression was performed to assess the impact of some confounding factors, namely conventional embryo morphology and female age. The associations between time-lapse embryo algorithms and treatment outcomes were expressed as an odds ratio (OR, 95% CI) before and after adjustment for confounding factors. Proportions were compared using the chi-squared test. All statistical analysis was performed with the use of the Statistic Package for the Social Sciences version 20.0 (SPSS; SPSS Inc., Chicago, IL) and $p < 0.05$ was considered statistically significant.

Results

Predictive power of time-lapse algorithms comparing KID and SET (implantation and live birth outcomes)

Table 3 shows the predictive power of the four time-lapse algorithms when applied to the same datasets. All four algorithms showed significant prediction power when applied to the KID dataset, expressed as AUC (95% CI): Liu 0.762 (0.701–0.824), KIDScore 0.614 (0.539–0.688), Meseguer 0.585 (0.508–0.663) and Basile 0.582 (0.505–0.659), $p < 0.05$, respectively. However, when the algorithms were applied to the SET dataset using implantation or live birth as the end-point, all showed a pattern of decreasing predictive power. The reduction in the AUC value was most noticeable

Table 3. ROC analysis of 4 time-lapse algorithms on KID and SET datasets using either implantation or live birth as the end-point.

	Predictive power of algorithms expressed as AUC (95% CI)		
	Implantation as end-point		Live birth as end-point
Algorithms	KID ($n = 270$)	SETs ($n = 144$)	SETs ($n = 144$)
Liu et al. (2016)	0.762 (0.701–0.824)*	0.724 (0.641–0.807)*	0.707 (0.620–0.793)*
KIDScore (Petersen et al., 2016)	0.614 (0.539–0.688)*	0.548 (0.451–0.645)	0.536 (0.434–0.637)
Meseguer et al. (2011)	0.585 (0.508–0.663)*	0.56 (0.462–0.658)	0.549 (0.445–0.652)
Basile et al. (2015)	0.582 (0.505–0.659)*	0.519 (0.421–0.618)	0.509 (0.406–0.612)

*$p < 0.05$.
KID: known implantation data; SET: single embryos transfer; AUC: area under the curve; CI: confidence interval.

between the KID and SET dataset using the same end-point (i.e. implantation) (Table 3). Despite the decrease in predictive power, the Liu et al. (2016) algorithm was the only algorithm that remained significantly predictive for either implantation or live birth. The other three algorithms lost their statistical significance of predictive power once they left the KID dataset (Table 3).

The impact of female age on implantation rates when embryo grading is similar

Table 4 compares the implantation rates of embryos with similar grading (judged via either conventional morphology or time-lapse features) but originating from different age groups of female patients. Between the two age groups (<35 years and ≥35 years), significant differences were detected when embryos were graded conventional score 1 (38.7 vs. 21.6%) or 2 (27.6 vs.7.7%), Liu score C (36.4% vs. 0/18), KIDScore 1 (38 vs. 20.2%) or 2 (20 vs. 2.5%), Meseguer score B+ (42.9 vs. 6.7%) or C+ (41.5 vs. 19.5%) and Basile score A+ (36.8 vs. 20%) or C+ (57.1 vs. 7.1%), $p < 0.05$, respectively.

Multiple regression analysis considering confounding factors

Table 5 shows the association between the four time-lapse algorithms and treatment outcomes before and after adjustment for the two selected potentially contributing factors – conventional day 3 embryo morphological score and female age. Analysis shows the Liu et al. (2016) algorithm remained significantly associated with treatment outcomes before and after adjustment for conventional embryo morphology with or without female age in all three of the datasets coupled with different end-points. The other three algorithms showed a statistically significant association with implantation in the KID dataset only prior to any adjustments – KIDScore (1.832, 1.118–3.004), Meseguer (1.150, 1.021–1.295) and Basile (1.122, 1.008–1.249).

Table 4. Implantation rates in the KID dataset according to female age and embryo grades.

	Female age groups	
	<35 years	≥35 years
Conventional morphology scores		
1	38.7% (36/93)*	21.6% (16/74)*
2	27.6% (8/29)*	7.7% (3/39)*
3	0/10	5.6% (1/18)
4	0/1	0/6
Liu et al. (2016)		
A+	61.9% (13/21)	38.5% (5/13)
A	40.8% (20/49)	29.4% (10/34)
B	20% (2/10)	30% (3/10)
C	36.4% (4/11)*	0/18*
D	26.7% (4/15)	5.9% (1/17)
E	3.8% (1/26)	2.6% (1/39)
F	0/1	0/6
KIDScore (Petersen et al., 2016)		
1	38% (38/100)*	20.2% (17/84)*
2	20% (5/25)*	2.5% (1/40)*
3	25% (1/4)	0/2
4	0/2	12.5% (1/8)
5	0/2	33.3% (1/3)
Meseguer et al. (2011)		
A+	40.6% (13/32)	29.2% (7/24)
A–	20% (2/10)	9.1% (1/11)
B+	42.9% (3/7)*	6.7% (1/15)*
B–	0/6	0/7
C+	41.5% (22/53)*	19.5% (8/41)*
C–	0/1	16.7% (1/6)
D+	23.1% (3/13)	6.3% (1/16)
D–	0/5	0/2
E	20% (1/5)	11.1% (1/9)
F	0/1	0/6
Basile et al. (2015)		
A+	36.8% (25/68)*	20% (12/60)*
A–	37.5% (3/8)	0/4
B+	0/8	18.2% (2/11)
B–	0/3	0/2
C+	57.1% (8/14)*	7.1% (1/14)*
C–	31.3% (5/16)	15% (3/20)
D+	66.7% (2/3)	33.3% (1/3)
D–	0/7	0/8
E	20% (1/5)	11.1% (1/9)
F	0/1	0/6

*Statistically significant difference between two age groups $p < 0.05$.
KID: known implantation data.

Once the algorithms were adjusted for conventional morphology, all three lost statistical significance (1.109, 0.694–1.772; 1.117, 0.980–1.273; 1.062, 0.945–1.194; respectively). This was also true when the datasets were adjusted for female age in-conjunction with morphology (1.063, 0.659–1.715; 1.122, 0.981–1.284;

Table 5. Associations between time-lapse algorithms and treatment outcomes following logistic regression analysis using different datasets (KID =270 vs. SET =144 with implantation and live birth as the end-point).

	Association between time-lapse algorithms and treatment outcomes		
	Odds ratio (95% CI)	Odds ratio adjusted for conventional morphology (95% CI)	Odds ratio adjusted for conventional morphology and female age (95% CI)
Known implantation data			
Liu et al. (2016)	1.837 (1.494–2.259)*	1.769 (1.394–2.244)*	1.746 (1.370–2.226)*
KIDScore (Petersen et al., 2016)	1.832 (1.118–3.004)*	1.109 (0.694–1.772)	1.063 (0.659–1.715)
Meseguer et al. (2011)	1.150 (1.021–1.295)*	1.117 (0.980–1.273)	1.122 (0.981–1.284)
Basile et al. (2015)	1.122 (1.008–1.249)*	1.062 (0.945–1.194)	1.038 (0.919–1.172)
SET with implantation as end-point			
Liu et al. (2016)	1.757 (1.353–2.281)*	1.743 (1.276–2.382)*	1.726 (1.259–2.365)*
KIDScore (Petersen et al., 2016)	1.257 (0.729–2.166)	0.865 (0.536–1.690)	0.892 (0.504–1.580)
Meseguer et al. (2011)	1.102 (0.945–1.286)	1.088 (0.926–1.278)	1.073 (0.910–1.264)
Basile et al. (2015)	1.022 (0.906–1.153)	0.984 (0.866–1.117)	0.973 (0.854–1.109)
SET with live birth as end-point			
Liu et al. (2016)	1.727 (1.307–2.281)*	1.572 (1.143–2.162)*	1.560 (1.133–2.147)*
KIDScore (Petersen et al., 2016)	1.118 (0.656–1.907)	0.772 (0.419–1.422)	0.740 (0.404–1.357)
Meseguer et al. (2011)	1.081 (0.920–1.270)	1.067 (0.899–1.266)	1.057 (0.890–1.256)
Basile et al. (2015)	1.011 (0.891–1.147)	0.964 (0.843–1.104)	0.958 (0.836–1.098)

*$p < 0.05$; SET: single embryos transfer; CI: confidence interval.

1.038, 0.919–1.172; respectively). When the KIDScore, Meseguer and Basile algorithms were applied to the SET datasets with either implantation or live birth as the end-point, none showed significant association with treatment outcomes before or after adjustment for conventional embryo morphology and female age (Table 5).

Discussion

Since the introduction of time-lapse technology into IVF laboratories, a number of factors have been found to influence embryo morphokinetics (Ciray, Aksoy, Goktas, Ozturk, & Bahceci, 2012; Freour, Dessolle, Lammers, Lattes, & Barriere, 2013; Kirkegaard, Hindkjaer, & Ingerslev, 2013; Munoz et al., 2012, 2013; Wale & Gardner, 2010). Such factors, originating from different patient or culture conditions between laboratories, are considered to contribute to the poor transferability of developed algorithms (Barrie et al., 2017; Best, Campbell, Duffy, Montgomery, & Fishel, 2013; Freour et al., 2015; Liu et al., 2015; Petersen et al., 2016; Yalcinkaya et al., 2014). The transferability issue is largely related to differences in the observed cleavage rates between embryos in different laboratories. To overcome this, a number of qualitative measures have been reported in recent years for embryo deselection (Athayde Wirka et al., 2014; Liu et al., 2014; Liu et al., 2015a; Rubio et al., 2012). These measures are categorized by abnormal cleavage patterns detected via time-lapse observation and are independent of absolute cleavage rates and timings of embryos. Such features could potentially improve inter-laboratory transferability of any proposed time-lapse algorithm (Liu et al., 2015a). In this study, 3 out of 4 (not KIDScore) algorithms use both qualitative and quantitative measures to define the final embryo score (grade), however, all have different emphasis upon the two types of measures (see Table 2).

There is limited published evidence showing positive external validation of the four models. Recently, the Liu et al. (2016) algorithm was positively favoured amongst several models using an external, large-scale, multi-centre blastulation dataset (Petersen et al., 2016). In this study, reduced predictive power was identified for three models (compared to the originally reported AUC as shown in Table 2) for KIDScore (AUC = 0.614, 0.539–0.688, $p < 0.05$), Meseguer algorithm (AUC = 0.585, 0.508–0.663, $p < 0.05$), and Basile algorithm (AUC = 0.582, 0.505–0.659, $p < 0.05$) (Table 3). This could potentially be related to different weightings given to the quantitative vs. qualitative measures in each of the models. Also, the dataset in the current model was originally used to develop the Liu et al. (2016) algorithm, and thus high predictive power of Liu model would be expected. Further external validation of the reported algorithms is warranted in future studies.

KID has been commonly used in the recent time-lapse studies, particularly when developing and/or validating an embryo selection algorithm in a dataset containing a significant number of double embryo transfer (DET) cycles (Basile et al., 2015; Liu et al., 2016; Meseguer et al., 2011; Petersen et al., 2016). The embryos in such studies were considered as independent observations with an end-point of either KID positive (+) or KID negative (−). KID + embryos refer to those originating from either a SET cycle where a single foetal heart is detected under ultrasound or a DET cycle where two foetal hearts are

detected. Whilst KID- embryos refer to those from a cycle (either a SET or DET) with a negative outcome regardless of the number of embryos transferred. However, the use of KID data excludes embryos from DET cycles with singleton pregnancy outcomes, owing to the impossibility of knowing which embryo implanted. A recent study raised concerns over the suitability of statistical methods used when analysing IVF/ICSI treatment outcomes, such as implantation or live birth, where embryos rather than patients were treated as independent observations (Kirkegaard et al., 2016). Sibling embryos in DETs are not independent subjects but are in fact related. In addition, DET is more likely to occur in poorer prognostic patients either with poorer quality embryos, or more advanced age, potentially leading to poorer treatment outcomes. Thus, to avoid confounding factors from the selection of DET, it may be more sensible to analyse individual patients (with one individual embryo), as the independent observation, when analysing treatment outcomes. Another potential issue of KID data is it can skew the dataset of treatment cycles due to the exclusion of positive cycles resulting in a single foetal heart (which could be the majority of positive cycles). Conversely, KID data also includes all negative cycles regardless of the number of embryos transferred. Therefore, the confounding effect of patient selection may bias the interpretation of an observed association between embryo grading and treatment outcome. In the SET dataset, however, embryo quality is considered one of the contributing factors to the treatment outcome, whilst treating individual patients as independent observations. Similarly, the exclusion of repeated treatment cycles by the same patients also ensures observations are independent of each other. As such, this study investigated different statistical outputs when analysing the predictive power of time-lapse algorithms on the treatment outcomes by using different datasets and different end-points. It is concerning to see the overestimated predictive power when using a KID dataset rather than SETs (with the exclusion of repeat cycles) to develop and validate time-lapse algorithms (Table 3). The reduction of predictive power on all the included algorithms when moving from a KID dataset to SETs is likely due to the removal of confounding variables (i.e. patient-related factors but not solely embryo), which would have been carried into the interpretation of statistical analysis. Therefore, it is paramount for future studies and clinics, when developing or validating time-lapse algorithms, to base conclusions on SETs rather than KID datasets, preferably using live birth as the end-point. A caveat to this though is not all IVF clinics have a large enough SET dataset for analysis, and many countries around the world still favour DET over SET.

Whilst appearing promising in a number of studies, time-lapse grading of human embryos should not be assumed to be an independent factor determining treatment outcome without further validation considering confounding factors. According to the data shown in Table 4, morphologically similar embryos may lead to different chance to implant in different female age groups. This may be owing to the different chromosomal or genetic constitution in embryos between younger and older patients, which may leave the morphology and morphokinetics of embryos unaffected. ROC analysis with AUC is a useful statistical tool when developing and validating diagnostic tests, and it has been employed in most of the recently published time-lapse embryo grading systems (Basile et al., 2015; Liu et al., 2016; Meseguer et al., 2011; Milewski et al., 2015; Petersen et al., 2016). However, ROC should be coupled with logistic regression analysis to verify the independence of observed association between proposed embryo grading and treatment outcome, by including other widely acknowledged factors, such as female age and conventional embryo morphology (Alpha Scientists in Reproductive Medicine & ESHRE Special Interest Group of Embryology, 2011; van Kooij, Looman, Habbema, Dorland, & te Velde, 1996). Results in this study have indicated the overriding power of conventional morphological score of embryos when simultaneously considering KIDScore, Meseguer, or Basile algorithm, respectively (Table 5). Therefore, caution must be employed when performing retrospective analysis, by including major potential confounders for better-controlled analysis.

One of the limitations of this study is its retrospective nature. Although a few selected confounding factors were considered during multiple regression analysis, other potential confounding effects from both known and unknown factors could not be eliminated in the current setting. Indeed, an RCT would have the capability to balance potential bias from diverse sources via randomization when comparing different groups, but such an approach is not always easily accessible (Kaser & Racowsky, 2014; Racowsky et al., 2015). Another limitation of the study is the use of the same dataset that was used for the development of the Liu et al. (2016) algorithm. This may have contributed to its superior prediction outcomes in the comparison. However, this study focused on the effect of: (a) choosing different datasets (e.g. KID vs. SET) and end-points (e.g.

implantation vs. live birth); and (b) inclusion of potential confounding factors, on the validity of conclusions drawn in the studies exploring time-lapse algorithms. It is recommended that the same analysis is replicated by an external organization to further validate the findings in this article. Additionally, the time of mixing sperm and oocytes in the conventional IVF is a procedural time point and does not represent any actual biological event for each individual oocyte within the same cohort. Furthermore, even in the ICSI cases, sperm entry time point for each individual oocyte would also be inaccurately recorded if only a single t0 is given to a whole cohort of oocytes. Therefore, in the Liu algorithm (Liu et al., 2016), tpnf was used to remove such uncertainty and inaccuracy, and also to unify IVF and ICSI embryos into a single model. However, in all the other three included algorithms, t0 was defined as either the sperm/oocyte mixing time or middle time point of the ICSI procedure. Therefore, when testing these models, the originally defined t0 was used. Finally, this study did not include all published time-lapse algorithms for comparison. The four algorithms were included based on the similarity of conditions which they were developed from, such as: (a) the use of retrospective KID dataset; (b) solely based on day 3 embryos; and (c) the use of the Embryoscope™ as the time-lapse incubator. It was anticipated that such an arrangement would improve comparability between the included algorithms. Indeed, other published algorithms from landmark studies, such as the Wong et al. (2010) algorithm were not included in this study due to: (a) being developed using blastocyst formation rather than implantation as an endpoint; (b) not being based on the Embryoscope system; and (c) parameters being generated from 2 rather than 3 d of culture with resulting inborn disadvantages for comparison. However, a recent large-scale retrospective study (Petersen et al., 2016) did include a number of other time-lapse algorithms including Eeva I & II, but clearly both did not do well.

In conclusion, SETs (with repeat patient cycles removed) are potentially a preferred type of dataset to develop or validate time-lapse embryo grading algorithms, ideally using live birth as an endpoint, to avoid potential overestimation of their predictive power. Multiple regression analysis is required to validate conclusions drawn through ROC analysis, by considering potential confounding factors contributing to the treatment outcome, such as conventional embryo morphology scores and female age.

Acknowledgements

The authors would like to thank the clinical and embryology teams at Fertility North for their technical support during this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology. (2011). The Istanbul consensus workshop on embryo assessment: Proceedings of an expert meeting. *Human Reproduction*, *26*, 1270–1283. doi: 10.1093/humrep/der037.

Athayde Wirka, K., Chen, A.A., Conaghan, J., Ivani, K., Gvakharia, M., Behr, B., … Shen, S. (2014). Atypical embryo phenotypes identified by time-lapse microscopy: High prevalence and association with embryo development. *Fertility and Sterility*, *101*, 1637–1648.e1–e5. doi: 10.1016/j.fertnstert.2014.02.050.

Barrie, A., Homburg, R., McDowell, G., Brown, J., Kingsland, C., & Troup, S. (2017). 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. *Fertility and Sterility*, *107*, 613–621. doi: 10.1016/j.fertnstert.2016.11.014.

Basile, N., Vime, P., Florensa, M., Aparicio Ruiz, B., Garcia Velasco, J.A., Remohi, J., & Meseguer, M. (2015). The use of morphokinetics as a predictor of implantation: A multicentric study to define and validate an algorithm for embryo selection. *Human Reproduction*, *30*, 276–283. doi: 10.1093/humrep/deu331.

Best, L., Campbell, A., Duffy, S., Montgomery, S., & Fishel, S. (2013). Does one model fit all? Testing a published embryo selection algorithm on independent timelapse data. *Human Reproduction*, 28*(1)*, i87–i90. doi: 10.1093/humrep/det190.

Ciray, H.N., Aksoy, T., Goktas, C., Ozturk, B., & Bahceci, M. (2012). Time-lapse evaluation of human embryo development in single versus sequential culture media–a sibling oocyte study. *Journal of Assisted Reproduction and Genetics*, *29*, 891–900. doi: 10.1007/s10815-012-9818-7.

Ciray, H.N., Campbell, A., Agerholm, I.E., Aguilar, J., Chamayou, S., Esbert, M., & Sayed, S. (2014). Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group. *Human Reproduction*, *29*, 2650–2660. doi: 10.1093/humrep/deu278.

Cohen, J., Alikani, M., & Bisignano, A. (2012). Past performance of assisted reproduction technologies as a model to predict future progress: A proposed addendum to Moore's law. *Reproductive Biomedicine Online*, *25*, 585–590. doi: 10.1016/j.rbmo.2012.08.006.

Conaghan, J., Chen, A.A., Willman, S.P., Ivani, K., Chenette, P.E., Boostanfar, R., … Shen, S. (2013). Improving embryo selection using a computer-automated time-lapse image analysis test plus day 3 morphology: Results from a

prospective multicenter trial. *Fertility and Sterility*, *100*, 412–419.e4–e15. doi: 10.1016/j.fertnstert.2013.04.021.

Dar, S., Lazer, T., Shah, P.S., & Librach, C.L. (2014). Neonatal outcomes among singleton births after blastocyst versus cleavage stage embryo transfer: A systematic review and meta-analysis. *Human Reproduction Update*, *20*, 439–448. doi: 10.1093/humupd/dmu001.

Freour, T., Dessolle, L., Lammers, J., Lattes, S., & Barriere, P. (2013). Comparison of embryo morphokinetics after in vitro fertilization-intracytoplasmic sperm injection in smoking and nonsmoking women. *Fertility and Sterility*, *99*, 1944–1950. doi: 10.1016/j.fertnstert.2013.01.136.

Freour, T., Le Fleuter, N., Lammers, J., Splingart, C., Reignier, A., & Barriere, P. (2015). External validation of a time-lapse prediction model. *Fertility and Sterility*, *103*, 917–922. doi: 10.1016/j.fertnstert.2014.12.111.

Gardner, D. (2016). The impact of physiological oxygen during culture, and vitrification for cryopreservation, on the outcome of extended culture in human IVF. *Reproductive Biomedicine Online*, *32*, 137–141. doi: 10.1016/j.rbmo.2015.11.008.

Gardner, D., & Balaban, B. (2016). Assessment of human embryo development using morphological criteria in an era of time-lapse, algorithms and 'OMICS': Is looking good still important? *Molecular Human Reproduction*, *22*, 704–718. doi: 10.1093/molehr/gaw057.

Gardner, D., Meseguer, M., Rubio, C., & Treff, N.R. (2015). Diagnosis of human preimplantation embryo viability. *Human Reproduction Update*, *21*, 727–747. doi: 10.1093/humupd/dmu064.

Gardner, D., Vella, P., Lane, M., Wagley, L., Schlenker, T., & Schoolcraft, W.B. (1998). Culture and transfer of human blastocysts increases implantation rates and reduces the need for multiple embryo transfers. *Fertility and Sterility*, *69*, 84–88. doi: 10.1016/S0015-0282(97)00438-X.

Kallen, B., Finnstrom, O., Lindam, A., Nilsson, E., Nygren, K.G., & Olausson, P.O. (2010). Blastocyst versus cleavage stage transfer in in vitro fertilization: Differences in neonatal outcome? *Fertility and Sterility*, *94*, 1680–1683. doi: 10.1016/j.fertnstert.2009.12.027.

Kalra, S.K., Ratcliffe, S.J., Barnhart, K.T., & Coutifaris, C. (2012). Extended embryo culture and an increased risk of preterm delivery. *Obstetrics & Gynecology*, *120*, 69–75. doi: 10.1097/AOG.0b013e31825b88fc.

Kaser, D.J., & Racowsky, C. (2014). Clinical outcomes following selection of human preimplantation embryos with time-lapse monitoring: A systematic review. *Human Reproduction Update*, *20*, 617–631. doi: 10.1093/humupd/dmu023.

Kirkegaard, K., Campbell, A., Agerholm, I., Bentin-Ley, U., Gabrielsen, A., Kirk, J., … Ingerslev, H.J. (2014). Limitations of a time-lapse blastocyst prediction model: A large multicentre outcome analysis. *Reproductive Biomedicine Online*, *29*, 156–158. doi: 10.1016/j.rbmo.2014.04.011.

Kirkegaard, K., Hindkjaer, J.J., & Ingerslev, H.J. (2013). Effect of oxygen concentration on human embryo development evaluated by time-lapse monitoring. *Fertility and Sterility*, *99*, 738–744.e4. doi: 10.1016/j.fertnstert.2012.11.028.

Kirkegaard, K., Sundvall, L., Erlandsen, M., Hindkjaer, J.J., Knudsen, U.B., & Ingerslev, H.J. (2016). Timing of human preimplantation embryonic development is confounded by embryo origin. *Human Reproduction*, *31*, 324–331. doi: 10.1093/humrep/dev296.

Liu, Y., Chapple, V., Feenan, K., Roberts, P., & Matson, P. (2015a). Clinical significance of intercellular contact at the four-cell stage of human embryos, and the use of abnormal cleavage patterns to identify embryos with low implantation potential: A time-lapse study. *Fertility and Sterility*, *103*, 1485–1491.e1. doi: 10.1016/j.fertnstert.2015.03.017.

Liu, Y., Chapple, V., Feenan, K., Roberts, P., & Matson, P. (2015b). Time-lapse videography of human embryos: Using pronuclear fading rather than insemination in IVF and ICSI cycles removes inconsistencies in time to reach early cleavage milestones. *Reproductive Biology*, *15*, 122–125. doi: 10.1016/j.repbio.2015.03.002.

Liu, Y., Chapple, V., Feenan, K., Roberts, P., & Matson, P. (2016). Time-lapse deselection model for human day 3 in vitro fertilization embryos: The combination of qualitative and quantitative measures of embryo growth. *Fertility and Sterility*, *105*, 656–662.e1. doi: 10.1016/j.fertnstert.2015.11.003.

Liu, Y., Chapple, V., Roberts, P., & Matson, P. (2014). Prevalence, consequence, and significance of reverse cleavage by human embryos viewed with the use of the Embryoscope time-lapse video system. *Fertility and Sterility*, *102*, 1295–1256. doi: 10.1016/j.fertnstert.2014.07.1235.

Liu, Y., Copeland, C., Stevens, A., Feenan, K., Chapple, V., Myssonski, K., … Matson, P. (2015). Assessment of human embryos by time-lapse videography: A comparison of quantitative and qualitative measures between two independent laboratories. *Reproductive Biology*, *15*, 210–216. doi: 10.1016/j.repbio.2015.09.001.

Maheshwari, A., Hamilton, M., & Bhattacharya, S. (2016). Should we be promoting embryo transfer at blastocyst stage? *Reproductive Biomedicine Online*, *32*, 142–146. doi: 10.1016/j.rbmo.2015.09.016.

Maheshwari, A., Kalampokas, T., Davidson, J., & Bhattacharya, S. (2013). Obstetric and perinatal outcomes in singleton pregnancies resulting from the transfer of blastocyst-stage versus cleavage-stage embryos generated through in vitro fertilization treatment: A systematic review and meta-analysis. *Fertility and Sterility*, *100*, 1615–1621. e1–e10. doi: 10.1016/j.fertnstert.2013.08.044.

Martins, W.P., Nastri, C.O., Rienzi, L., van der Poel, S.Z., Gracia, C., & Racowsky, C. (2016). Blastocyst vs cleavage-stage embryo transfer: Systematic review and meta-analysis of reproductive outcomes. *Ultrasound in Obstetrics & Gynecology*, *49*, 583–591. doi: 10.1002/uog.17327.

Meseguer, M., Herrero, J., Tejera, A., Hilligsoe, K.M., Ramsing, N.B., & Remohi, J. (2011). The use of morphokinetics as a predictor of embryo implantation. *Human Reproduction (Oxford, England)*, *26*, 2658–2671. doi: 10.1093/humrep/der256.

Milewski, R., Kuc, P., Kuczynska, A., Stankiewicz, B., Lukaszuk, K., & Kuczynski, W. (2015). A predictive model for blastocyst formation based on morphokinetic parameters in time-lapse monitoring of embryo development. *Journal of Assisted Reproduction and Genetics*, *32*, 571–579. doi: 10.1007/s10815-015-0440-3.

Montag, M., Toth, B., & Strowitzki, T. (2013). New approaches to embryo selection. *Reproductive Biomedicine Online*, *27*, 539–546. doi: 10.1016/j.rbmo.2013.05.013.

Munoz, M., Cruz, M., Humaidan, P., Garrido, N., Perez-Cano, I., & Meseguer, M. (2012). Dose of recombinant FSH and oestradiol concentration on day of HCG affect embryo development kinetics. *Reproductive Biomedicine Online*, *25*, 382–389. doi: 10.1016/j.rbmo.2012.06.016.

Munoz, M., Cruz, M., Humaidan, P., Garrido, N., Perez-Cano, I., & Meseguer, M. (2013). The type of GnRH analogue used during controlled ovarian stimulation influences early embryo developmental kinetics: A time-lapse study. *European Journal of Obstetrics, Gynecology and Reproductive Biology*, *168*, 167–172. doi: 10.1016/j.ejogrb.2012.12.038.

Petersen, B.M., Boel, M., Montag, M., & Gardner, D.K. (2016). Development of a generally applicable morphokinetic algorithm capable of predicting the implantation potential of embryos transferred on Day 3. *Human Reproduction*, *31*, 2231–2244. doi: 10.1093/humrep/dew188.

Racowsky, C., Kovacs, P., & Martins, W.P. (2015). A critical appraisal of time-lapse imaging for embryo selection: Where are we and where do we need to go? *Journal of Assisted Reproduction and Genetics*, *32*, 1025–1030. doi: 10.1007/s10815-015-0510-6.

Rubio, I., Galan, A., Larreategui, Z., Ayerdi, F., Bellver, J., Herrero, J., & Meseguer, M. (2014). Clinical validation of embryo culture and selection by morphokinetic analysis: A randomized, controlled trial of the EmbryoScope. *Fertility and Sterility*, *102*, 1287–1294.e5. doi: 10.1016/j.fertnstert.2014.07.738.

Rubio, I., Kuhlmann, R., Agerholm, I., Kirk, J., Herrero, J., Escriba, M.J., … Meseguer, M. (2012). Limited implantation success of direct-cleaved human zygotes: A time-lapse study. *Fertility and Sterility*, *98*, 1458–1463. doi: 10.1016/j.fertnstert.2012.07.1135.

van Kooij, R.J., Looman, C.W., Habbema, J.D., Dorland, M., & Te Velde, E.R. (1996). Age-dependent decrease in embryo implantation rate after in vitro fertilization. *Fertility and Sterility*, *66*, 769–775. doi: 10.1016/S0015-0282(16)58634-8.

Wale, P.L., & Gardner, D.K. (2010). Time-lapse analysis of mouse embryo development in oxygen gradients. *Reproductive Biomedicine Online*, *21*, 402–410. doi: 10.1016/j.rbmo.2010.04.028.

Wong, C.C., Loewke, K.E., Bossert, N.L., Behr, B., De Jonge, C.J., Baer, T.M., & Reijo Pera, R.A. (2010). Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nature Biotechnology*, *28*, 1115–1121. doi: 10.1038/nbt.1686.

Yalcinkaya, E., Ergin, E.G., Caliskan, E., Oztel, Z., Ozay, A., & Ozornek, H. (2014). Reproducibility of a time-lapse embryo selection model based on morphokinetic data in a sequential culture media setting. *Journal of the Turkish German Gynecological Association*, *15*, 156–160. doi: 10.5152/jtgga.2014.13068.