Human Reproduction, Vol.32, No.2 pp. 307–314, 2017
Advanced Access publication on December 28, 2016 doi:10.1093/humrep/dew330

human reproduction

ORIGINAL ARTICLE *Embryology*

Inter-observer and intra-observer agreement between embryologists during selection of a single Day 5 embryo for transfer: a multicenter study

Ashleigh Storr[1,2,*], Christos A. Venetis[1,2], Simon Cooke[1,2], Suha Kilani[1,2], and William Ledger[1,2]

[1]IVFAustralia, Sydney, NSW, Australia [2]School of Women's and Children's Health, UNSW Medicine, UNSW, NSW, Australia

*Correspondence address. IVFAustralia Eastern Suburbs—Maroubra, 1/225 Maroubra Rd, Maroubra, NSW 2035, Australia. Fax: +61 (0) 2 9349 5888; E-mail: ashleigh.storr@outlook.com

Submitted on February 16, 2016; resubmitted on November 21, 2016; accepted on December 7, 2016

STUDY QUESTION: What is the inter-observer and intra-observer agreement between embryologists when selecting a single Day 5 embryo for transfer?

SUMMARY ANSWER: The inter-observer and intra-observer agreement between embryologists when selecting a single Day 5 embryo for transfer was generally good, although not optimal, even among experienced embryologists.

WHAT IS KNOWN ALREADY: Previous research on the morphological assessment of early stage (two pronuclei to Day 3) embryos has shown varying levels of inter-observer and intra-observer agreement. However, single blastocyst transfer is now becoming increasingly popular and there are no published data that assess inter-observer and intra-observer agreement when selecting a single embryo for Day 5 transfer.

STUDY DESIGN, SIZE, DURATION: This was a prospective study involving 10 embryologists working at five different IVF clinics within a single organization between July 2013 and November 2015.

PARTICIPANTS/MATERIALS, SETTING, METHODS: The top 10 embryologists were selected based on their yearly Quality Assurance Program scores for blastocyst grading and were asked to morphologically grade all Day 5 embryos and choose a single embryo for transfer in a survey of 100 cases using 2D images. A total of 1000 decisions were therefore assessed. For each case, Day 5 images were shown, followed by a Day 3 and Day 5 image of the same embryo. Subgroup analyses were also performed based on the following characteristics of embryologists: the level of clinical embryology experience in the laboratory; amount of research experience; number of days per week spent grading embryos. The agreement between these embryologists and the one that scored the embryos on the actual day of transfer was also evaluated. Inter-observer and intra-observer variability was assessed using the kappa coefficient to evaluate the extent of agreement.

MAIN RESULTS AND THE ROLE OF CHANCE: This study showed that all 10 embryologists agreed on the embryo chosen for transfer in 50 out of 100 cases. In 93 out of 100 cases, at least 6 out of the 10 embryologists agreed. The inter-observer and intra-observer agreement among embryologists when selecting a single Day 5 embryo for transfer was generally good as assessed by the kappa scores (kappa = 0.734, 95% CI: 0.665–0.791 and 0.759, 95% CI: 0.622–0.833, respectively). The subgroup analyses did not substantially alter the inter-observer and intra-observer agreement among embryologists. The agreement when Day 3 images were included alongside Day 5 images of the same embryos resulted in a change of mind at least three times by each embryologist (on average for <10% of cases) and resulted in a small decrease in inter-observer and intra-observer agreement between embryologists (kappa = 0.676, 95% CI: 0.617–0.724 and 0.752, 95% CI: 0.656–808, respectively).

The assessment of the inter-observer agreement with regard to morphological grading of Day 5 embryos showed only a fair-to-moderate agreement, which was observed across all subgroup analyses. The highest overall kappa coefficient was seen for the grading of the

developmental stage of an embryo (0.513; 95% CI: 0.492–0.538). The findings were similar when the individual embryologists were compared with the embryologist who made the morphological assessments of the available embryos on the actual day of transfer.

LIMITATIONS, REASONS FOR CAUTION: All embryologists had already completed their training and were working under one organization with similar policies between the five clinics. Therefore, the inter-observer agreement might not be as high between embryologists working in clinics with different policies or with different levels of training.

WIDER IMPLICATIONS OF THE FINDINGS: The generally good, although not optimal uniformity between participating embryologists when selecting a Day 5 embryo for transfer, as well as, the surprisingly low agreement when morphologically grading Day 5 embryos could be improved, potentially resulting in increased pregnancy rates. Future studies need to be directed toward technologies that can help achieve this.

STUDY FUNDING/COMPETING INTEREST(S): None declared.

TRIAL REGISTRATION NUMBER: Not applicable.

Key words: agreement / embryo / inter-observer variability / intra-observer variability / morphology

Introduction

The outcome of IVF is dependent on a combination of patient characteristics, clinical management and laboratory practice. Ultimately, one of the major determining factors for pregnancy achievement after IVF is embryo quality (Gardner *et al.*, 2000; Ahlstrom *et al.*, 2011), and, for this reason, several grading systems have been developed in order to accurately identify the embryo with the highest implantation potential (Steer *et al.*, 1992; Gardner *et al.*, 2000; Fisch *et al.*, 2001; Depa-Martynow *et al.*, 2007).

Although these grading systems are quite detailed and use a number of morphological parameters, this process is still highly subjective because it involves the interpretation and application of these criteria by individual embryologists (Arce *et al.*, 2006; Baxter Bendus *et al.*, 2006; Ahlstrom *et al.*, 2011). Some variation in embryo scoring between different assessors is therefore to be expected, even within the same clinic, however, this may compromise the accuracy of morphology-based embryo selection (Arce *et al.*, 2006).

This becomes most relevant in the context of single embryo transfer, since the embryo with the highest implantation potential has to be selected for transfer. Extending embryo culture to the blastocyst stage provides more information about embryo development and quality and allows the embryologist to make a more informed selection from a smaller cohort of embryos (Van Royen *et al.*, 1999; Sjoblom *et al.*, 2006; Ahlstrom *et al.*, 2011). This optimizes pregnancy rates while avoiding the risks associated with multiple pregnancies (Gardner *et al.*, 1998; Van der Auwera *et al.*, 2002; Mangalraj *et al.*, 2009). For these reasons, single blastocyst transfer (SBT) is gradually becoming the preferable strategy in many clinics worldwide (Glujovsky *et al.*, 2012).

Previous research on the morphological assessment of early stage (two pronuclei (2PN) to Day 3) embryos has shown varying levels of inter-observer and intra-observer agreement (Arce *et al.*, 2006; Baxter Bendus *et al.*, 2006; Paternot *et al.*, 2009; Paternot *et al.*, 2011). The evaluation of Day 5 embryos is significantly different to that of early stage embryos, as a blastocyst structure is more complex and introduces more variables to consider when deciding on the appropriate embryo for transfer (Gardner *et al.*, 2000). Whether this has an impact on the ability of embryologists to identify the same embryo as the most suitable for transfer is not currently known. If the inter-observer and intra-observer agreement for embryologists in selecting the best embryo for transfer is poor, this may lead to inconsistent and less than optimal pregnancy rates.

For this reason, the aim of the current study was to evaluate the inter-observer and intra-observer agreement of 10 embryologists from different clinics in selecting the best Day 5 embryo (the one with presumably the highest implantation potential) for transfer, as well as the inter-observer agreement between these embryologists when morphologically grading Day 5 embryos.

Materials and Methods

Study design

This was a prospective study including 10 embryologists working at five different clinics within IVFAustralia in New South Wales between July 2013 and November 2015. The aim was to assess the inter-observer and intra-observer agreement in selecting the best Day 5 embryo (the one with presumably the highest implantation potential) for transfer and also evaluate the inter-observer agreement in grading (using standard morphology) Day 5 embryos.

Survey design and study participants

Cases for this study (n = 100) were chosen from a database of EmbryoViewer (Vitrolife, Aarhus, Denmark) images collected from previously performed cycles in a single center where the EmbryoScope (Vitrolife, Aarhus, Denmark) is available (Storr *et al.*, 2015). The included cases were from a preclinical phase where the EmbryoScope was used as a standard incubator. An independent embryologist selected these cases based on the development of more than one potential embryo for transfer on Day 5, resulting in a total of 428 embryos analyzed. Each embryo had a single 2D Day 5 image copied from the EmbryoViewer (340 × 340 mm; 72 pixels/inch) taken immediately before embryo transfer, which was specific to each patient. A Day 3 image of the same embryo at 76 hours post insemination was also copied. Images were chosen by an independent embryologist to best represent the morphology of the embryo at that time point by adjusting the focus and light exposure of the images. If any of the embryos could not be visualized correctly, the case was not included in the study. Using the copied images, a questionnaire was developed with a web survey designer (SurveyMonkey Inc., Palo Alto, CA, USA) where each image was made distinguishable to the researchers (but not to the participating embryologists) using a unique identification code.

The 10 highest scoring IVFAustralia embryologists were selected based on the 2014 annual results obtained from their monthly online Quality Assurance Program (QAP) for blastocyst grading (FertAid, Newcastle, New South Wales, Australia). Each embryologist had undergone 12–18 months of in-house training by all fully trained members of their team as well as completed 2 years of supervised blastocyst quality control as a trainee before they could undertake embryo assessment unaccompanied at IVFAustralia. The participants' characteristics are summarized in Table I. Most of the participating embryologists had research experience and four had a Master's degree or higher. All participants were involved in embryo grading at least 2 days per week. The participants were blinded to the unique identification codes of the cases as well as to the evaluations made by other participants. Images in the survey were presented to the participants on a case-by-case basis, first asking which embryo they would select for transfer based on Day 5 images (Question 1), and then asking the same question providing Day 3 and Day 5 images side-by-side (Question 2). The participants were asked to only select the best embryo for transfer as if they were in a laboratory setting. The participants were allowed to complete the surveys separately and in more than one sitting. Six months after the initial survey, the same participants completed the same survey again and additionally were asked to provide information on Day 5 embryo quality. If a participating embryologist considered an embryo to be a full blastocyst, an expanded blastocyst or a hatching blastocyst, then they were also asked to provide a grade for inner cell mass (ICM) and trophectoderm. If an embryologist considered an embryo to be a morula, a very early blastocyst or an early blastocyst, then they were asked to provide a quality grade (A or B) for that embryo (Gardner *et al.*, 2000). Blastocysts with ICM A or B and trophectoderm grade A (i.e. AA or BA) were considered to be top quality (Ahlstrom *et al.*, 2011), blastocysts with AB or BB grading were considered to be good quality, while blastocysts with AC, BC, CA, CB or CC grading were considered to be poor quality. Morulas and early blastocysts with A grade were considered good quality embryos, whereas B grade morulas and early blastocysts were considered to be poor quality embryos. The second survey was performed in order to assess intra-observer agreement between the first and second surveys, as well as to assess inter-observer agreement in the various aspects of morphological grading. Participating embryologists were not aware that they would be asked to repeat the survey a second time, and they did not have access to the data collected in the first survey.

Sample size

Since there were no published data available regarding inter-observer and intra-observer variability on Day 5 embryo selection and this study aimed at comparing inter-observer agreement between more than two assessors, a sample size analysis was not feasible. For this reason, the study was designed as exploratory and it was considered that 1000 decisions (10 embryologists deciding on 100 cases) would be both adequate and realistic.

Outcomes

The primary outcome of this study was to determine the inter-observer and intra-observer agreement between embryologists in selection of a blastocyst for transfer based on images of Day 5 embryos. Moreover, this study aimed to establish whether inter-observer and intra-observer agreement for Day 5 embryo selection were affected by addition of Day 3 morphology. Subgroup analyses were also performed based on the following characteristics of embryologists: level of experience in the laboratory, amount of research experience and number of days per week grading embryos. Furthermore, it was assessed whether the majority of surveyed embryologists (>5) agreed with the embryo that had been selected on the day of transfer as well as whether the quality of embryos available in a cohort on Day 5 (e.g. all embryos of poor quality or >1 top quality embryo) affected agreement between embryologists.

Table I Characteristics of the participants in a study of agreement between embryologists during selection of a single Day 5 embryo for transfer.

Participant number	Age group (years)	Gender	Qualifications	Research experience	Professional experience	Position	Days per week at work	Days per week grading embryos	Site
1	41–45	Female	Bachelor of Science	None	>10 years	Supervisor	4	2	2
2	36–40	Female	Bachelor of Science	Up to 1 year	>10 years	Senior embryologist	5	2	3
3	31–35	Female	MSc or equivalent	Up to 1 year	5–10 years	Embryologist	5	>3	4
4	36–40	Male	PhD or higher	3–5 years	5–10 years	Supervisor	5	2	1
5	31–35	Female	Bachelor of Science	None	>10 years	Senior embryologist	5	2	5
6	26–30	Female	MSc or equivalent	1–3 years	3–5 years	Embryologist	5	3	3
7	31–35	Female	Bachelor of Science	None	>10 years	Senior embryologist	3	3	4
8	26–30	Female	MSc or equivalent	Up to 1 year	3–5 years	Embryologist	5	3	1
9	41–45	Female	Bachelor of Science	None	>10 years	Senior embryologist	4	3	3
10	46–50	Female	Bachelor of Science	1–3 years	>10 years	Senior embryologist	5	2	4

The inter-observer agreement for Day 5 embryo morphological grading was also analyzed to determine which aspect of embryo grading was the most challenging to gain agreement on. The inter-observer agreement between the individual embryologists and the embryologist who selected the embryo on the day of transfer in the laboratory was also examined. Furthermore, the inter-observer agreement for the morphological assessment of the embryos made by the individual embryologists and the embryologist on the day of transfer was also determined.

Although this study was not powered to evaluate pregnancy outcomes, for reasons of comprehensiveness, pregnancy rates based on the quality of the embryo transferred are reported, as well as the pregnancy rates when the embryo selected by the majority of embryologists (>5) had been actually transferred.

Statistical analysis

The Fleiss kappa coefficient (k) (Fleiss, 1971) was used in order to evaluate the extent of agreement. The kappa coefficient is a chance corrected index that can measure the agreement between selections made by different raters (Landis and Koch, 1977). The maximum k equals to 1 and a k equal to 0 represents no agreement between raters. For intermediate values, an interpretation that has been suggested is <0.20: poor; 0.21–0.40: fair; 0.41–0.60: moderate; 0.61–0.80: good and 0.81–1.00: very good (Altman, 1990).

The 95% CI of the kappa was produced using bootstrapping (Efron and Tibshirani, 1994). More specifically, 100 repetitions were drawn for the calculation of each CI. All statistical analyses were performed with STATA (StataCorp., v.14.1, TX, USA). Statistical significance was set at $P \leq 0.05$.

Table II Summary of Day 5 embryo characteristics.

Characteristic	Total cases
Total number of embryos	428
Mean number of embryos per case (SD)	4.28 (1.99)
Number of embryos per case	Number of cases
2	14
3	27
4	21
≥5	38
All poor quality embryos	3
More than one top quality embryo	27

Results

The 10 embryologists included in this study (Table I) represent ~25% of the total number of embryologists working at IVFAustralia. A summary of the characteristics of embryos included in this analysis can be found in Table II.

Inter-observer agreement

Decision based on Day 5 assessment

Table III displays the overall agreement between the embryologists involved in the survey when asked to decide on which embryo to transfer based on Day 5 images alone (Question 1). All 10 embryologists agreed on the embryo chosen for transfer in 50 out of 100 cases. In 93 out of 100 cases, at least 6 out of the 10 embryologists agreed.

The kappa scores reflecting the level of agreement are displayed in Table IV. The kappa score for all embryologists (0.734, 95% CI: 0.665–0.791) represents good agreement. The level of experience as an embryologist, level of research experience and number of days per week grading embryos did not substantially affect the kappa agreement coefficient (Supplementary Table SI).

When there was more than one top quality embryo to choose from in a single case, the kappa score (0.745, 95% CI: 0.667–0.871) was similar to that seen when there was one or no top quality embryos to choose from (0.725, 95% CI: 0.666–0.795). When all embryos were of poor quality in a single case, the kappa score (0.718, 95% CI: 0.504–0.841) was similar to that seen when there was at least one other embryo of better quality in the cohort (0.734, 95% CI: 0.648–0.775).

The kappa score for the inter-observer agreement between the embryo that had been transferred in the laboratory and the embryo selected by the majority (>5) of participating embryologists for Day 5 images only (Question 1) was 0.752, 95% CI: 0.643–0.868 (agreement: 81.1%).

Decision based on Day 3 and Day 5 assessment

Table III displays the overall agreement between the embryologists involved in the survey when asked to decide on which embryo to transfer based on Day 3 and Day 5 images together (Question 2). All 10 embryologists agreed on the embryo chosen for transfer in 41 out of 100 cases. At least 6 out of the 10 embryologists agreed on the same embryo for transfer in 88 out of 100 cases.

Table III Extent of agreement between embryologists regarding which Day 5 embryo to select for transfer.

Agreement	Question 1 % (95% CI)	Cumulative % (95% CI)	Question 2 % (95% CI)	Cumulative % (95% CI)	Total % (95% CI)	Cumulative % (95% CI)
10/10	50 (40.4–59.6)	50 (40.4–59.6)	41 (31.9–51.8)	41 (31.9–51.8)	46 (39.2–52.9)	46 (39.2–52.9)
9/10	15 (9.3–23.3)	65 (55.3–73.6)	19 (12.5–27.8)	60 (50.2–69.1)	17 (12.4–22.8)	63 (56.1–69.4)
8/10	10 (5.5–17.4)	75 (65.7–82.5)	8 (4.1–15.0)	68 (58.3–76.3)	9 (5.8–13.8)	72 (65.4–77.8)
7/10	7 (3.4–13.7)	82 (73.3–88.3)	13 (7.8–21.0)	81 (72.2–87.5)	10 (6.6–14.9)	82 (76.1–86.7)
6/10	11 (6.3–18.6)	93 (86.3–96.6)	7 (3.4–13.7)	88 (80.2–93.0)	9 (5.8–13.8)	91 (86.2–94.2)

Question 1 includes images for Day 5 embryos only. Question 2 includes images for Day 3 and Day 5 embryos. Total refers to the number of agreements for both Questions 1 and 2 ($n = 200$).

Table IV Kappa scores for the inter-observer and intra-observer agreement of all embryologists in the selection of an embryo for transfer based on Day 5 images alone (Question 1) as well as Day 3 and Day 5 images together (Question 2).

Agreement	Question 1*	Question 2*	Overall*
Inter-observer agreement	0.734 (0.665–0.791)	0.676 (0.617–0.724)	0.705 (0.669–0.756)
Intra-observer agreement	0.759 (0.622–0.833)	0.752 (0.656–0.808)	0.753 (0.661–0.820)

*Kappa scores with 95% CI presented.

The kappa scores reflecting the level of agreement for this question (Question 2) are displayed in Table IV. The kappa score for all embryologists (0.676, 95% CI: 0.617–0.724) represents good agreement. The level of experience as an embryologist, level of research experience and number of days per week grading embryos also did not substantially affect the kappa agreement coefficient, although a small but consistent drop was observed from Question 1 to Question 2 (Supplementary Table SI). The mean number of times the embryologists changed their answer from Question 1 to Question 2 (based on the additional Day 3 images provided) was 8.4 (SD 4.2), ranging from 3 to 15. When the embryologists changed their original decision, then in 20 cases (for all embryologists, i.e. mean: two decisions per embryologist), the embryo they selected was the same as the one actually transferred. In 53 decisions (mean: 5.3 decisions per embryologist), the embryologists changed their original selection from an embryo that agreed with the one actually selected for transfer in the laboratory to a different one.

When there was more than one top quality embryo to choose from in a single case, the kappa score (0.674, 95% CI: 0.574–0.802) was similar to that seen when there was one or no top quality embryos to choose from (0.672, 95% CI: 0.608–0.767). When all embryos were of poor quality in a single case, the kappa score (0.637, 95% CI: 0.420–0.841) was similar to that seen when there was at least one other embryo of better quality in the cohort (0.677, 95% CI: 0.633–0.747).

The kappa score for the inter-observer agreement between the embryo that had been transferred in the laboratory and the embryo selected by the majority (>5) of embryologists for Day 3 and Day 5 images (Question 2) was 0.700, 95% CI: 0.578–0.812 (agreement: 77.7%).

Morphological grading

A fair kappa score was observed for the overall agreement between the embryologists when grading ICM (0.349, 95% CI: 0.301–0.392) and trophectoderm (0.397, 95% CI: 0.356–0.423) for embryos with full blastulation (full blastocyst, expanded blastocyst and hatching blastocyst). For embryos without full blastulation (morula, very early blastocyst and early blastocyst), the overall agreement on quality (A or B) was also fair (kappa = 0.393, 95% CI: 0.340–0.472). The highest kappa coefficient was observed when grading the developmental stage of an embryo (i.e. whether an embryo was judged as a morula, very early blastocyst, early blastocyst, full blastocyst, expanded blastocyst or hatching blastocyst) (0.513; 95% CI: 0.492–0.538), representing a moderate agreement (Table V).

The grading agreement between the embryologists based on ICM, trophectoderm, quality and developmental stage according to the subgroup analyses was also assessed. The level of experience as an embryologist, level of research experience and number of days per week grading embryos did not substantially alter the kappa agreement coefficient (Supplementary Table SII).

Table V Kappa scores for the inter-observer agreement between embryologists according to ICM, trophectoderm, quality and developmental stage.

Embryo morphology grading	Kappa (95% CI)
ICM[a]	0.349 (0.301–0.392)
Trophectoderm[b]	0.397 (0.356–0.423)
Quality[c]	0.393 (0.340–0.472)
Developmental stage[d]	0.513 (0.492–0.538)

[a]Quality grading of ICM of fully blastulated embryos (i.e. full blastocyst, expanded blastocyst and hatching blastocyst) as judged by each participating embryologist.
[b]Quality grading of trophectoderm of fully blastulated embryos (i.e. full blastocyst, expanded blastocyst and hatching blastocyst) as judged by each participating embryologist.
[c]Quality grading of embryos prior to blastulation (i.e. morula, very early blastocyst and early blastocyst) as judged by each participating embryologist.
[d]Developmental stage of an embryo without quality grading, as judged by each participating embryologist (i.e. whether an embryo was judged as a morula, very early blastocyst, early blastocyst, full blastocyst, expanded blastocyst or hatching blastocyst).

The kappa scores for the inter-observer agreement between the Day 5 embryo selected for transfer by the individual study embryologists and that selected by the embryologist on the actual day of transfer ranged from 0.569 (95% CI: 0.477–0.662) to 0.674 (95% CI: 0.576–0.757) (Supplementary Table SIII). When the agreement between each individual study embryologist and the embryologist that made the morphological grading for ICM on the actual day of transfer was compared, the kappa scores ranged from −0.032 (95% CI: −0.063 to −0.009) to −0.010 (95% CI: −0.092 to 0.083). When the trophectoderm grading was analyzed, the kappa scores ranged from −0.114 (95% CI: −0.152 to −0.077) to 0.110 (95% CI: 0.038–0.211). When the quality of the embryo was compared, the kappa scores ranged from 0.166 (95% CI: 0.038–0.294) to 0.422 (95% CI: 0.276–0.567). The kappa scores for the grading of developmental stage ranged from 0.299 (95% CI: 0.247–0.353) to 0.408 (95% CI: 0.327–0.446) (Supplementary Table SIII).

Intra-observer agreement

Decision based on Day 5 assessment

The kappa scores for intra-observer agreement of the embryologists when asked to decide on which embryo to transfer based on Day 5 images alone (Question 1) ranged from 0.662 (95% CI: 0.524–0.759) to 0.833 (95% CI: 0.734–0.920) (Supplementary Table SIV).

The kappa scores reflecting the level of agreement for Question 1 are displayed in Table IV. The kappa score for all embryologists (0.759, 95% CI: 0.622–0.833) represents good agreement. The level of experience as an embryologist, level of research experience and number of days per week grading embryos did not substantially affect the kappa agreement coefficient (Supplementary Table SV).

Decision based on Day 3 and Day 5 assessment

The kappa scores for the intra-observer agreement between the embryologists when asked to decide on which embryo to transfer based on Day 3 and Day 5 images together (Question 2) ranged from 0.656 (95% CI: 0.568–0.797) to 0.808 (95% CI: 0.729–0.903) (Supplementary Table SIV).

The kappa scores reflecting the level of intra-observer agreement for Question 2 are displayed in Table IV. The kappa score for all embryologists (0.752, 95% CI: 0.656–0.808) represents good agreement. The level of experience as an embryologist, level of research experience and number of days per week grading embryos did not substantially affect the kappa agreement coefficient (Supplementary Table SV).

Pregnancy outcomes

If the embryo chosen by the majority of embryologists was the same as that which had been chosen for transfer in the laboratory, the clinical pregnancy rate was 30.3% (20/66). If the embryo chosen by the majority of embryologists was different to that which had been chosen for transfer in the laboratory, the clinical pregnancy rate was 31.6% (6/19) (odds ratio: 0.942, 95% CI: 0.313–2.832; $P = 0.915$). There were 15 cases for which there was no embryo selected by the majority (>5) of participating embryologists.

The pregnancy rates depending on the quality of the embryo transferred as judged in the laboratory were as follows: poor embryo: 16.7% (95% CI: 0.0–46.4) (1/6), good embryo: 27.5% (95% CI: 13.7–41.3) (11/40) and top embryo: 53.7% (95% CI: 40.4–67.0) (29/54).

Discussion

This study showed that the inter-observer and intra-observer agreement among embryologists when selecting a single embryo for transfer on Day 5 is generally good, although not optimal.

These findings were relatively stable in a variety of different subgroup analyses based on the characteristics of the embryologists or on the cohort of the examined embryos. More specifically in the current study, the subgroup analyses based on the level of experience as an embryologist, the level of research experience and the number of days per week grading embryos did not lead to an increased inter-observer or intra-observer agreement of participants. However, it should be noted that the participating embryologists had all completed their in-house training for blastocyst assessment and were the 10 highest scoring embryologists based on their 2014 annual QAP. It might be hypothesized that embryologists at earlier stages of training could show lower levels of agreement.

This was suggested in a study by Paternot *et al.* (2009), where inter-observer agreement was higher overall among experienced embryologists than among trainees when five embryologists assessed multiple quality parameters of 50 embryos on Days 1, 2 and 3 (Paternot *et al.*, 2009). On the other hand, a study by Baxter Bendus *et al.* (2006) did not include trainees in their analysis of 35 Day 3 embryos, and found no effect of experience or level of education on the inter-observer and intra-observer agreement for multiple parameters of embryo quality (Baxter Bendus *et al.*, 2006). Despite the fact that these studies were conducted to assess early embryo scoring instead of choosing an embryo for transfer on Day 5, their results, when taken together with the findings of the current study, might mean that fully trained embryologists have good agreement overall in the assessments they perform in the laboratory.

This seems to also be supported in a more recent study by Paternot *et al.* (2011) where good inter-observer agreement (kappa = 0.71, 95% CI 0.60–0.86) and intra-observer agreement (kappa = 0.75, 95% CI 0.72–0.88) were noted when assessing five embryologists on deciding the final outcome of 180 Day 1, Day 2 and Day 3 embryos, i.e. if it should be transferred, cryopreserved or discarded (Paternot *et al.*, 2011). Similarly, Arce *et al.* (2006) found good-to-excellent inter-observer (33 local and 3 central) and intra-observer agreement (3 central) of embryologists when assessing Days 1, 2 and 3 embryo quality parameters of 4002 embryos (Arce *et al.*, 2006).

The present study also assessed the agreement when Day 3 images were included alongside Day 5 images of the same embryo. The additional information resulted in a change of mind at least three times for each embryologist (on average <10%) and caused a small decrease in agreement between embryologists, which also persisted in every subgroup analyses performed. This finding is unsurprising, since provision of additional information adds more complexity to the decision making process and increased diversity in the conclusions reached. In line with this is the fact that the agreement of the surveyed embryologists on which embryo to transfer with the decision made in the laboratory on the day of transfer was decreased when Day 3 images were included alongside Day 5 images of the same embryo. It should be noted that the decision made in the laboratory took into account the embryo grade on Day 3 noted on the embryology worksheet since at that stage the EmbryoScope was used as a standard incubator. Moreover, a different embryologist potentially made the assessment of this grade and this could further explain the aforementioned decrease in agreement.

In the current study, the embryo quality within the cohort (whether all embryos were of poor quality or whether two or more top quality embryos were present) did not substantially alter the inter-observer agreement between embryologists, as assessed by the kappa scores. This is an interesting finding considering that the presence of more than one top quality embryo within a cohort would be expected to make it more challenging to select the single best embryo for transfer.

The assessment of the inter-observer agreement with regard to the grading of ICM, trophectoderm, quality and developmental stage of embryos revealed that the agreement between participating embryologists was not good across the different aspects of embryo grading. Furthermore, a higher agreement was also seen between embryologists when they were asked to select an embryo for transfer, rather than provide a morphology grade. This outcome was also observed for the agreement between the Day 5 embryo selected for transfer by the individual study embryologists and that selected by the embryologist on the actual day of transfer, which was higher than the same comparison made for the morphological grading of embryos. A potential explanation might be that it is easier to select the best embryo out of a cohort of embryos, whereas providing the same morphology grade for the

individual characteristics of an embryo (ICM, trophectoderm or developmental stage) might be more challenging, resulting in a lower agreement. Something that also needs to be taken into account regarding the corresponding values of the kappa coefficient is that the kappa coefficient is a chance corrected index. Consequently, for the same percentage agreement, the kappa coefficient will be lower for a question in which the number of potential answers is lower than for a question in which the number of potential answers is higher (Bakeman *et al.*, 1997).

Previous studies that have assessed the agreement between embryologists when grading early stage (2PN to Day 3) embryos have observed good-to-excellent agreement for the scoring of blastomeres (number and size) (Arce *et al.*, 2006; Paternot *et al.*, 2009, 2011) and multinucleation on Day 2 of embryo development (Arce *et al.*, 2006). Other parameters assessed range from poor-to-moderate agreement, with the agreement decreasing as the embryo develops (Arce *et al.*, 2006; Paternot *et al.*, 2009, 2011). In the current study, only a fair-to-moderate agreement was seen for the morphological grading of Day 5 embryos, supporting the suggestion made in previous studies that as an embryo becomes more complex with development, the agreement between embryologists declines (Arce *et al.*, 2006; Paternot *et al.*, 2009, 2011).

One of the strengths of the current study is its large size. A total of 10 embryologists assessed Day 5 embryos in 100 cases resulting in 1000 decisions made, more than in the previously published studies assessing inter-observer and intra-observer variability (Baxter Bendus *et al.*, 2006; Paternot *et al.*, 2009, 2011). Additionally, the research questions addressed in this study are relevant to the decisions occurring every day in the laboratory during morphology-based embryo selection, and since Day 5 single embryo transfer is also becoming more widespread (Glujovsky *et al.*, 2012), the outcomes of this study are applicable to current practice in the laboratory. Finally, the level of agreement between embryologists was assessed using the Fleiss kappa coefficient (Fleiss, 1971). This is generally considered to be a robust measure for assessing agreement between >2 rates on a nominal scale as it takes into account the fact that the agreement might have occurred by chance (Landis and Koch, 1977).

This study has some limitations that need to be discussed. First, the number of embryologists involved in the selection of embryos for transfer was smaller than some of the previously published studies (Arce *et al.*, 2006; Baxter Bendus *et al.*, 2006). However, the total number of cases was not small, especially considering that this was a study assessing embryo selection in the context of SBT (2000 decisions made for Question 1 and Question 2 combined) (Baxter Bendus *et al.*, 2006; Paternot *et al.*, 2009, 2011). While the current study was performed across five different clinics, each one is part of a single organization operating under similar policies and the inter-observer agreement might not be as high between embryologists working in clinics with different policies. Finally, the use of 2D images does not allow embryologists to move or examine embryos at a higher magnification, which can occur when choosing an embryo for transfer in the laboratory. However, every effort was made to select an image that was the best representation of embryo quality on Day 3 and Day 5 and if embryos were not able to be visualized correctly, the case was not included in the study.

The kappa coefficients calculated in this study in regard to embryo selection and morphology grading show a less than optimal level of inter-observer and intra-observer agreement for the participating embryologists across all aspects of analysis, and theoretically could be improved (Paternot *et al.*, 2009). A higher level of consistency when grading and choosing an embryo for transfer may result in the improvement in pregnancy rates after SBT and offer the best chance of success for every patient. Although increased training for embryologists may help improve the level of agreement (Paternot *et al.*, 2009), it should be noted that the embryologists involved in this study were highly trained in blastocyst assessment. Hence, it may be that more advanced technologies, such as time-lapse based algorithms, could allow the more consistent selection of the best embryo for transfer by removal of the human factor altogether. This type of technology is still maturing (Armstrong *et al.*, 2015; Kirkegaard *et al.*, 2015), but there is potential that future research will produce validated tools that will standardize embryo selection and optimize pregnancy rates. At present, the results of this study suggest that current embryo selection and grading methods can be further improved and perhaps more emphasis on the validation of time-lapse technology is required to allow implementation of such tools in the future.

In conclusion, the inter-observer and intra-observer agreement between embryologists when selecting a single Day 5 embryo for transfer was generally good, although not optimal. Furthermore, there was a lack of uniformity when grading individual morphological characteristics of Day 5 embryos, even among experienced embryologists. Future studies need to be directed toward tools that can increase the consistency in embryo grading and selection.

Supplementary data

Supplementary data are available at *Human Reproduction* online.

Acknowledgements

The authors wish to thank the 10 embryologists at IVFAustralia for their continued support throughout this project, as well their extensive involvement in the survey for this analysis.

Authors' roles

A.S. conceived the idea of the study, contributed in the construction of the protocol, performed the analyses and interpretation of the data and drafted the manuscript. C.A.V. conceived the idea of the study, contributed in the construction of the protocol, performed the analyses and interpretation of the data and revised the manuscript for important intellectual content. S.C., S.K. and W.L. contributed in the interpretation of the data and revised the manuscript for important intellectual content.

Funding

No funding was received for this study.

Conflict of interest

None declared.

References

Ahlstrom A, Westin C, Reismer E, Wikland M, Hardarson T. Trophectoderm morphology: an important parameter for predicting live birth after single blastocyst transfer. *Hum Reprod* 2011;**26**:3289–3296.

Altman DG. *Practical Statistics for Medical Research*. London: Chapman Hall/CRC Press, 1990.

Arce JC, Ziebe S, Lundin K, Janssens R, Helmgaard L, Sorensen P. Interobserver agreement and intraobserver reproducibility of embryo quality assessments. *Hum Reprod* 2006;**21**:2141–2148.

Armstrong S, Arroll N, Cree LM, Jordan V, Farquhar C. Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database Syst Rev* 2015;**2**:CD011320.

Bakeman R, McArthur D, Quera V, Robinson BF. Detecting sequential patterns and determining their reliability with fallible observers. *Psychol Methods* 1997;**2**:357.

Baxter Bendus AE, Mayer JF, Shipley SK, Catherino WH. Interobserver and intraobserver variation in day 3 embryo grading. *Fertil Steril* 2006;**86**: 1608–1615.

Depa-Martynow M, Jedrzejczak P, Pawelczyk L. Pronuclear scoring as a predictor of embryo quality in in vitro fertilization program. *Folia Histochem Cytobiol* 2007;**45**:S85–89.

Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton FL: Chapman & Hall/CRC Press, 1994.

Fisch JD, Rodriguez H, Ross R, Overby G, Sher G. The graduated embryo score (GES) predicts blastocyst formation and pregnancy rate from cleavage-stage embryos. *Hum Reprod* 2001;**16**:1970–1975.

Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;**76**:378.

Gardner DK, Lane M, Stevens J, Schlenker T, Schoolcraft WB. Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertil Steril* 2000;**73**:1155–1158.

Gardner DK, Vella P, Lane M, Wagley L, Schlenker T, Schoolcraft WB. Culture and transfer of human blastocysts increases implantation rates and reduces the need for multiple embryo transfers. *Fertil Steril* 1998;**69**:84–88.

Glujovsky D, Blake D, Farquhar C, Bardach A. Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology. *Cochrane Database Syst Rev* 2012;**7**:CD002118.

Kirkegaard K, Ahlstrom A, Ingerslev HJ, Hardarson T. Choosing the best embryo by time lapse versus standard morphology. *Fertil Steril* 2015; **103**:323–332.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–174.

Mangalraj AM, Muthukumar K, Aleyamma T, Kamath MS, George K. Blastocyst stage transfer vs cleavage stage embryo transfer. *J Hum Reprod Sci* 2009;**2**:23–26.

Paternot G, Devroe J, Debrock S, D'Hooghe TM, Spiessens C. Intra- and inter-observer analysis in the morphological assessment of early-stage embryos. *Reprod Biol Endocrinol* 2009;**7**:105.

Paternot G, Wetzels AM, Thonon F, Vansteenbrugge A, Willemen D, Devroe J, Debrock S, D'Hooghe TM, Spiessens C. Intra- and interobserver analysis in the morphological assessment of early stage embryos during an IVF procedure: a multicentre study. *Reprod Biol Endocrinol* 2011;**9**:127.

Sjoblom P, Menezes J, Cummins L, Mathiyalagan B, Costello MF. Prediction of embryo developmental potential and pregnancy based on early stage morphological characteristics. *Fertil Steril* 2006;**86**:848–861.

Steer CV, Mills CL, Tan SL, Campbell S, Edwards RG. The cumulative embryo score: a predictive embryo scoring technique to select the optimal number of embryos to transfer in an in-vitro fertilization and embryo transfer programme. *Hum Reprod* 1992;**7**:117–119.

Storr A, Venetis CA, Cooke S, Susetio D, Kilani S, Ledger W. Morphokinetic parameters using time-lapse technology and day 5 embryo quality: a prospective cohort study. *J Assist Reprod Genet* 2015;**7**: 1151–1160.

Van der Auwera I, Debrock S, Spiessens C, Afschrift H, Bakelants E, Meuleman C, Meeuwis L, D'Hooghe TM. A prospective randomized study: day 2 versus day 5 embryo transfer. *Hum Reprod* 2002;**17**:1507–1512.

Van Royen E, Mangelschots K, De Neubourg D, Valkenburg M, Van de Meerssche M, Ryckaert G, Eestermans W, Gerris J. Characterization of a top quality embryo, a step towards single-embryo transfer. *Hum Reprod* 1999;**14**:2345–2349.