Reducing inter-observer variability in embryo evaluation by means of training courses

Rafael Ruiz de Assin · Ana Clavero · Maria Carmen Gonzalvo · Antonio Rosales · Sandra Zamora · Luis Martínez · Juan Mozas · Jose Antonio Castilla

Received: 12 April 2011 / Accepted: 12 September 2011 / Published online: 23 September 2011
© Springer Science+Business Media, LLC 2011

Abstract

Purpose To study the utility of a training session offered to junior embryologists, comparing the results obtained with those reported by a group of senior embryologists.

Methods The 62 junior embryologists participating were asked to decide on the quality of the embryos and the clinical decision to be taken.

Results The junior embryologists' success rate following the training course was significantly higher than before for

embryo classification ($48.4\% \pm 20.4$ vs. $59.7\% \pm 16.7$) ($p < 0.05$) and for clinical decision ($54.7\% \pm 19.6$ vs. $68.7\% \pm 17.6$) ($p < 0.005$). Comparison of the degree of agreement between the categories assigned by the junior embryologists and those assigned by consensus among the group of senior embryologists revealed kappa values of $k=0.32$ before the course and of $k=0.54$ after it. The comparison between pre- and post-training junior and senior embryologists also reflected an improvement in the kappa index for clinical decision, from $k=0.54$ to $k=0.68$.

Conclusions Training courses are shown to be an effective tool for increasing the degree of agreement between junior and senior embryologists.

Capsule Training courses are shown to be an effective tool for increasing the degree of agreement between junior and senior embryologists.

R. Ruiz de Assin · A. Clavero · M. C. Gonzalvo · A. Rosales · S. Zamora · L. Martínez · J. Mozas · J. A. Castilla
Unidad de Reproducción Humana,
Hospital Universitario "Virgen de las Nieves",
Granada, Spain

J. A. Castilla
Banco de Semen CEIFER,
Granada, Spain

J. A. Castilla
Programa de Control de Calidad Externo para el Laboratorio de Reproducción de la Asociación para el Estudio de la Biología de la Reproducción (ASEBIR),
Madrid, Spain

J. Mozas
Departamento de Obstetricia y Ginecología,
Universidad de Granada,
Granada, Spain

J. Mozas (✉)
Unidad de Reproducción,
C/ Dr. Azpitarte. Edificio de consultas externas.
Hospital Materno-Infantil,
Granada 18014, Spain
e-mail: jmozasm@sego.es

Keywords Embryo evaluation · Consensus session · Training courses

Introduction

The implantation ability of an embryo, and thus the possibility of pregnancy being achieved, depends on various factors, one of which is embryo morphology [1–4], and therefore evaluation of the latter is a key element in IVF/ICSI treatment. Among the factors that may affect the evaluation of embryo morphology is that of the different systems used for embryo classification and intra- and inter-observer variability [5–7].

The Spanish Association for the Study of Reproductive Biology (ASEBIR) has implemented various approaches to reduce the high variability in embryo evaluation described by different authors [6–10]. Thus, an External Quality Control Programme was initiated in 2003 [9]. Participation in such programmes for embryo evaluation is recommended by diverse scientific societies [11, 12], and it has been

shown that inclusion in this type of programme reduces differences between laboratories, regarding both embryo classification and the decision concerning the action to be taken [9, 10, 13]. Subsequently, in 2008 a guideline was published on the morphologic evaluation of oocytes, early-stage embryos and human blastocysts [11], as the lack of standardization in evaluation criteria is another of the major problems facing the embryologist when a decision must be taken as to whether an embryo is of good or poor quality. Some authors have championed the embryo scoring system [1–4, 14], while others prefer classification by categories [1, 7].

Another possible strategy for reducing inter-observer variability is that offered by training courses, which have proven to be useful in semen analysis [15, 16].

The aim of the present study was to investigate the utility of a training course for junior embryologists, comparing the results obtained with the evaluation performed by a group of senior embryologists, who had previously participated in a consensus session.

Material and methods

A study was made of the variability between junior embryologists before and after their participation in an ASEBIR training course. For the purposes of this study, videos of embryos at different stages of development (zygotes, Day 2 embryos and Day 3 embryos) were used, divided into four blocks with five embryos in each. The training sessions, each lasting one day, were performed at three sites in Spain (Barcelona, Madrid and Sevilla).

The junior embryologists were not randomly selected, but were invited to volunteer for participation in the project, and the only condition stipulated was that they should have less than 2 years' clinical experience.

Firstly, the junior embryologists carried out a pre-course test, via the Web, in which they were asked to evaluate two blocks of five embryos (one of Day 2 embryos and one of Day 3 embryos). In each case, they determined, individually, the quality of each embryo, as A (top quality), B (good quality), C (average quality) or D (poor quality). In addition, they were asked about the clinical decision to be taken regarding the embryos, assuming that each block corresponded to a different puncture and that the couple wished two embryos to be transferred; for the remaining embryos, the decision to be taken was whether they should be frozen or discarded.

After the training course, a post-course test was performed, via the Web, in which two blocks of five embryos (one of Day 2 embryos and one of Day 3 embryos) were evaluated. The videos for the post-course test were different from those of the pre-course test.

The answers given by the junior embryologists were compared with those provided by a group of senior embryologists who had previously taken part in a consensus session on embryo evaluation. To determine the utility of these consensus sessions and to study the effect they produced on inter-observer variability among a group of five senior embryologists, 140 videos were shown, of embryos at different stages of development (zygotes, Day 2 embryos and Day 3 embryos), divided into 28 blocks of five videos.

The five senior embryologists were selected on the basis of their expert knowledge of human embryology. They each have over 10 years' experience; all are members of the ASEBIR working group on embryo quality and four of them hold the title "Certified Senior Clinical Embryologist" by ESHRE (2008). The meeting, which took place in Madrid at the Gregorio Marañón Hospital, was divided into three parts. Firstly, a pre-consensus test was performed, in which the senior embryologists were asked to evaluate five blocks, each with five videos of embryos at different stages of division. Block 1 showed five videos of zygotes, Blocks 2 and 3 each contained five videos with Day 2 embryos, and Blocks 4 and 5 each corresponded to five videos showing Day 3 embryos. For each block, the senior embryologists were asked to categorize, individually, the quality of each of the embryos (as Good, Average or Poor) and to decide on the clinical decision to be taken for each one, assuming that each block corresponded to a different puncture and that the couple wished to keep only two zygotes in the culture medium (in the case of the zygotes) and transfer only two embryos (in the case of the Day 2 and Day 3 embryos). The senior embryologists were then asked whether the remaining zygotes and the Day 2 and Day 3 embryos should be frozen or discarded.

The second part of the meeting consisted of a consensus session, at which the senior embryologists were first shown the results obtained during the pre-consensus test, together with the corresponding videos, and these were discussed. Subsequently, the evaluation of 19 blocks of five embryos (95 embryos in total) was included. These 95 embryos were jointly evaluated by the senior embryologists, and the quality of each one in turn was discussed; for each of the 19 blocks, the decision was taken as to which two embryos should be preserved in the culture medium (in the case of the zygotes) or transferred (in the case of the Day 2 and Day 3 embryos), and for the other three embryos in each block, the senior embryologists decided which should be frozen and which should be discarded. During this phase of the meeting, the senior embryologists were allowed to consult the recommendations of the ASEBIR Clinical Embryology Guidelines, Volume II [12].

Finally, the third part of the meeting consisted of a post-consensus test, in which four blocks of five embryos were

evaluated (two blocks of Day 2 embryos and two blocks of Day 3 embryos). The videos of these 20 embryos were not viewed during the pre-consensus test or during the consensus session, so that neither the classification nor the clinical decision taken should be affected.

Table 1 shows the different activities carried out by the senior and junior embryologists.

In the case of the senior embryologists, there was considered to be agreement regarding the classification or clinical decision on an embryo when the five participants made the same choice, and there was considered to be disagreement when one or more of them disagreed with the others.

The χ^2 test, with a significance level of 5%, was used to compare the percentage of embryos for which there was observed to be agreement among the senior embryologists, before and after the consensus session, and the Student *t* test was used to compare the mean success rate of the junior embryologists, before and after the training course. Kappa reliability values were used to analyze the agreement between the categories assigned by the senior embryologists and those nominated by the majority of the training course participants, before and after the training course. To interpret the level of agreement measured by the kappa values of reliability, the proposal made by Landis and Koch [17] was utilized (≥ 0.80 : almost perfect agreement; 0.61 – 0.80 : substantial agreement; 0.41 – 0.60 : moderate agreement; 0.21 – 0.40 : fair agreement; ≤ 0.20 : slight agreement).

Results

Consensus session (senior embryologists)

Of the 20 embryos evaluated by the senior embryologists during the test performed following the consensus session, one (a Day 3 specimen) could not be evaluated, due to the poor quality of the video shown—in the participants' opinion—and so it was excluded from the results analysis.

Table 2 shows the embryo classification for the pre-consensus test compared with that for the post-consensus test. The differences are significant ($p < 0.05$) with respect to the percentage of embryos (35% vs. 73.7%), excluding the

Table 2 Percentage of embryos excluding the data corresponding to zygotes with agreement between senior embryologists before and after the consensus session

	Pre-consensus	Post-consensus	P
Classification	35.0% (7/20)	73.7% (14/19)	<0.05
Clinical Decision	30.0% (6/20)	84.2% (16/19)	<0.005

data corresponding to zygotes, for which agreement was reached before and after the consensus session.

In the pre-consensus test, agreement was reached on 9/25 (36.0%) of the embryos, of which 2/5 (40.0%) were at the zygote stage, 4/10 (40.0%) were Day 2 embryos, and 3/10 (30.0%) were Day 3 embryos. Of the nine videos on which there was agreement, three of the embryos were classified as Good quality, three as Average and three as Poor. In the post-consensus test, there was agreement in 14/19 cases (73.7%), of which 6/10 (60.0%) were Day 2 embryos and 8/9 (88.9%) were Day 3 embryos. Of the 14 videos on which there was agreement, eight of the embryos were classified as Good quality, five as Average and one as Poor.

With respect to the clinical decision on embryos excluding the data corresponding to zygotes, the differences are significant ($p < 0.005$) regarding agreement before and after the consensus session (30% vs. 84.2%).

In the pre-consensus test, there was agreement on 9/25 (36.0%) of the embryos, of which 3/5 (60.0%) were at the zygote stage, 4/10 (40.0%) were Day 2 embryos and 2/10 (20.0%) were Day 3 embryos. In the post-session test, there was agreement in 16/19 (84.2%) cases, of which 7/10 (70.0%) were Day 2 embryos and 9/9 (100%) were Day 3 embryos.

Training session (junior embryologists)

The success rate among the junior embryologists following the training course was significantly higher than before it ($48.4\% \pm 20.4$ vs. $59.7\% \pm 16.7$) ($p < 0.05$) (Fig. 1a).

On comparison of the agreement among the categories assigned by the majority of these junior embryologists with those determined by consensus among the senior embryologists, the kappa values were found to be $k = 0.32$ before the course and $k = 0.54$ after it.

With respect to the clinical decision, there was also seen to be a large improvement in the percentage of embryos on which the junior embryologists, following the training course, agreed with the senior embryologists ($54.7\% \pm 19.6$ vs. $68.7\% \pm 17.6$) ($p < 0.005$) (Fig. 1b).

We also observed a rise in the kappa index regarding the clinical decision adopted by the majority of the trainees, on completion of the course, and that of the senior embryologists, from $k = 0.54$ in the pre-course test to $k = 0.68$ after it.

Table 1 Activity carried out by senior and junior embryologists

	Senior	Junior
Test	Pre-consensus test	Pre-course test
Session	Consensus session	Training course
Test	Post-consensus test	Post-course test

Fig. 1 Percentage of embryos agreed upon by all of the junior embryologists in the training course. **a** Embryo classification; **b** Clinical decision

Discussion

There is general consensus that didactic lectures and discussion play an important role in improving the knowledge base, but the effective teaching of clinical skills requires more direct teacher-student interaction. It is obviously impossible to provide live embryos in a structured teaching situation with opportunities for interaction, but models can be designed that are very close to a real-life situation, for example via the use of videos of embryos at different stages of development.

Accordingly, following the publication of the ASEBIR Guideline for the morphologic evaluation of oocytes, early-stage embryos and human blastocysts [12], this Association proposed that courses should be organized with a high level of teacher-student interaction, and with this idea in mind, designed a series of appropriate training courses.

Our results show that established strategies for reducing inter-observer variability, such as participation in an external quality control programme [9] and the standardization of embryo classification criteria [12] would be enhanced by the addition of training courses. This finding coincides with the conclusions of studies on semen evaluation [15, 16].

In evaluating the utility of a training course, we decided to use as a reference value the opinions expressed as the consensus of senior embryologists, rather than the view of the majority of participants, because our research group has shown, previously, that the level of agreement among laboratories is lower than among senior embryologists, and that the agreement among senior embryologists and laboratories is only moderate [10]. Therefore, it is recommended that an assigned value in training courses should be established on the basis of consensus values obtained from senior embryologists.

There was found to be a higher degree of agreement among senior embryologists and junior embryologists on clinical decision than on embryo classification, which we believe is a very useful finding, as it is the clinical decision that is the crucial factor determining the technique

subsequently applied. We believe this result is due to the fact that it is easier to assimilate concepts that involve a comparison between embryos (better, worse) than more abstract concepts requiring individual evaluation for the embryos to be assigned to one of the various categories (top, good, average or poor).

There was found to be a better kappa index between the junior and senior embryologists following the training session, rising from a “fair” degree of agreement to “moderate” regarding embryo classification and from one of “moderate” to “substantial” agreement regarding the clinical decision.

This study does present several limitations, such as the use of video, for which only limited recording time was available, and in which the embryos could not be viewed from different angles, thus creating an artificial situation over which the embryologist had no control. Nevertheless, Arce et al. [6] have demonstrated the validity of a digital imaging system similar to ours that was used to compare the skills of embryologists. Another problem is that the method used for enrolling the junior embryologists in this course could have provoked the inclusion of some degree of bias. The junior embryologists were not randomly selected, but were invited to volunteer for participation in the project, and the only condition stipulated was that they should have less than 2 years’ clinical experience. This voluntary participation might reflect selection bias, in which those junior embryologists who were particularly interested in the training course might have specifically opted to participate, whereas others, with less interest, might have chosen not to. Alternatively, the contrary might be the case, in which those junior embryologists with a less advanced background, and being aware of their poorer skills, could have decided to take part in this course in order to improve. Moreover, we are unaware of the influence of the dynamics and organization of the training course on the assimilation of its content, as similar experiments in the healthcare field have revealed differences in this respect, depending on the resources employed during the course in question [18, 19].

Finally, we do not know whether the observed increase in the degree of agreement among the five senior embryologists following the consensus session, and between this group and the junior embryologists after the training course, will persist over time, or the frequency with which it would be necessary to offer supplementary sessions and courses, in order to reduce the differences among embryologists in evaluating embryo quality. Prior studies carried out by our group [9, 13], have shown that participation in external quality control programmes that include the evaluation of embryos by the use of video-based instruction does tend to increase the degree of agreement among laboratories.

Recently, Dessolle et al. [20] demonstrated the importance of the cumulative summation test for the learning curve (LC-CUSUM) in teaching embryology laboratory techniques. CUSUM methodology allows real-time monitoring of key outcomes. Moreover, individual performance can be quantified and compared within a team or between laboratories. The application of these statistical methods in embryo evaluation and selection studies could be of considerable utility.

Whilst recognizing the above limitations, we have shown in the present study that when junior embryologists attended a training course in embryo evaluation, there were smaller differences between their evaluations and those of senior embryologists, regarding both the categories assigned and the clinical decision.

Acknowledgements The authors of this study thank the members of the embryo quality working group of the Spanish Association for the Study of Reproductive Biology (ASEBIR)—Manuel Ardoy (U Reproducción, Hospital Gregorio Marañón, Madrid), Jorge Cuadros (FIV Madrid, Madrid), María José Torelló (Clínica Quirón, Barcelona), Gema Arroyo (IU Dexeus, Barcelona) and Luz Rodríguez (Fundación Jiménez Díaz, Madrid)—for the confidence placed in the authors and for their cooperation with this work, without which this study could not have been made. In addition, we thank the coordinators of the training courses held in Barcelona (Montse Boada and Mark Grossman), Madrid (Jorge Cuadros and Manuel Ardoy) and Sevilla (Victoria Hurtado de Mendoza).

References

1. Sharpe-Timms KL, Zimmer RL. Oocyte and pre-embryo classification. In: Kal BA, May JV, De Jonge CI, editors. Handbook of the assisted reproduction laboratory. 1st ed. USA: CRC; 2000. p. 179–96.
2. 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;16:1970–5.
3. De Placido G, Wilding M, Strina I, Alviggi E, Alviggi C, Mollo A, et al. High outcome predictability alter IVF using a combined store for zygote and embryo morphology and growth rate. *Hum Reprod*. 2002;17:2402–9.
4. Holte J, Berglund L, Milton K, Garello C, Gennarelli G, Revelli A, et al. Construction of an evidence-based integrated morphology cleavage embryo score for implantation potential of embryos scored and transferred on day 2 after oocyte retrieval. *Hum Reprod*. 2007;22:548–57.
5. Keck C, Fischer R, Baukloh V, Alper M. Quality management in reproductive medicine. In: Gadner DK, Weissman A, Howles CM, Shohan Z, editors. Textbook of assisted reproductive techniques. Laboratory and clinical perspectives. 2nd ed. London and New York: Taylor and Francis; 2004. p. 477–94.
6. Arce JC, Ziebe S, Lundin K, Janssens R, Helmgaard L, Sorensen P. Interobserver agreement and intraobserver reproducibility of embryo quality assessments. *Hum Reprod*. 2006;21:2141–8.
7. Baxter AE, Mayer JF, Shipley SK, Catherino WH. Interobserver and intraobserver variation in day 3 embryo grading. *Fertil Steril*. 2006;86:1608–15.
8. Matson PL. Internal and external quality assurance in the IVF laboratory. *Hum Reprod*. 1998;13:156–65.
9. Castilla JA, Ruiz de Assin R, Gonzalvo MC, Clavero A, Ramirez JP, Vergara F, et al. External quality control for embryology laboratory. *RBMOnline*. 2010;20:68–74.
10. Ruiz de Assin R, Clavero A, Gonzalvo MC, Ramirez JP, Zamora S, Fernández A, et al. Comparison of methods to determine the assigned value in an external quality control programme for embryo evaluation. *RBM Online*. 2009;19:824–9.
11. The Practice Committee of the American Society for Reproductive Medicine and the Practice Committee of the Society for Assisted Reproductive Technology. Revised guidelines for human embryology and andrology laboratories. *Fertil Steril*. 2006;86(Suppl 4):57–72.
12. ASEBIR. Criterios de valoración morfológicos de oocitos, embriones tempranos y blastocistos humanos. II Cuaderno de Embriología Clínica. 2008 (www.asebir.com/.../cuadernos.../ii-cuadernos-de-embriologia-clinica).
13. Castilla JA, Ortiz A, Magán R, Ortiz-Galisteo JR, González E, Aguilar J, et al. Resultados de un ensayo piloto para un Programa Nacional de Control de Calidad Externo de Laboratorio de FIV. *ASEBIR*. 2003;8:40–5.
14. Desai NN, Goldstein J, Rowland DY, Goldfarb JM. Morphological evaluation of human embryos and derivation of an embryo quality scoring system specific for day 3 embryos: a preliminary study. *Hum Reprod*. 2000;15:2190–6.
15. Björndahl L, Barratt CLR, Fraser LR, Kvist U, Mortimer D. ESHRE basic semen analysis courses 1995–1999: immediate beneficial effects of standardized training. *Hum Reprod*. 2002;17:1299–305.
16. Franken DR, Kruger TF. Lessons learned from a sperm morphology quality control programme. *Andrologia*. 2006;38:225–9.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–74.
18. Ali J, Adam R, Pierre I, Bedaysie H, Josa D, Winn J. Comparison of performance 2 years after the old and new (interactive) ATLS courses. *J Surg Res*. 2001;97:71–5.
19. Ali J, Adam RU, Sammy I, Ali E, Williams JI. The simulated trauma patient teaching module—does it improve student performance? *J Trauma*. 2007;62:1416–20.
20. Dessolle L, Biau DJ, Larouzière V, Ravel C, Antoine J-M, Deraï E, et al. Learning curve of vitrification assessed by cumulative summation test for learning curve (LC-CUSUM). *Fertil Steril*. 2009;92:943–5.