Article

Comparison of methods to determine the assigned value in an external quality control programme for embryo evaluation

Dr Ruiz de Assin received his Diploma in biology in 2004 from the University of Granada, Spain. He received his Master's degree in 2007 from the University of Murcia for a study on biology and technology in mammal reproduction. He is currently a PhD student and embryologist in the Human Reproduction Unit at the Virgen de las Nieves University Hospital, Granada, Spain. His thesis focuses on external quality control programmes in IVF laboratories.

Dr Ruiz de Assin

R Ruiz de Assín[1,4], A Clavero[1], M C Gonzalvo[1], J P Ramírez[2,3], S Zamora[1], A Fernández[1], L Martínez[1], J A Castilla[1,2,3]
[1]Human Reproduction Unit, Virgen de las Nieves University Hospital, E-18014 Granada, Spain; [2]Sperm Bank CEIFER, Granada, Spain; [3]External Quality Control Programme for a Human Reproduction Laboratory, Association for the Study of Reproductive Biology (ASEBIR), Madrid, Spain
[4]Correspondence: e-mail: rafaraa@hotmail.com

Abstract

This study set out to establish adequate assigned values for a National External Quality Control Programme of embryo evaluation. The results obtained by Spanish laboratories in this programme are compared with those of a group of national experts in embryo quality. Image-based embryo evaluation consists not only of classifying embryos as being of optimal, moderate or poor quality, but also of specifying the clinical decision to be taken regarding each embryo (transfer, cryopreservation or rejection). The proportion of embryos for which there was a high degree of agreement among the experts was 98.3% for embryo classification and 93.3% for clinical decision; for the laboratories, the respective values were 44.2 and 42.5%. With respect to the interobserver agreement among laboratories and experts, kappa coefficients were lower than 0.6 both for classification and for clinical decision. The experts recommended cryopreservation of a higher percentage of embryos classified as poor quality than did the laboratories (28 versus 4%, $P = 0.05$). The data obtained show that the agreement among laboratories is lower than among experts, and that the concordance among experts and laboratories is moderate. Therefore, it is recommended that an assigned value from external quality control programmes is established based on the consensus values obtained from experts.

***Keywords**: assigned value, embryo quality, external quality control, interobserver*

Introduction

Participation in External Quality Control Programmes (EQCP) in embryo evaluation is recommended by various scientific societies (Practice Committee of the American Society for Reproductive Medicine and Practice Committee of the Society for Assisted Reproductive Technology, 2006; ASEBIR, 2008; Magli *et al.*, 2008). In Spain, since 2003, the Spanish Association for the Study of Reproductive Biology (ASEBIR) has promoted an EQCP of this type by sending a DVD/CD-ROM with videos of zygotes and of day 2 and day 3 embryos to participating laboratories; it has been found that the inclusion of such centres in this kind of programme increases the degree of inter-laboratory agreement in embryo classification (Castilla *et al.*, 2009).

On the other hand, the generalized lack of standardization concerning evaluation criteria is one of the main problems facing embryologists in deciding whether an embryo is of optimal or poor quality. Some authors have proposed systems based on embryo scoring (Desai *et al.*, 2000; Sharpe-Timms and Zimmer, 2000; Fisch *et al.*, 2001; De Placido *et al.*, 2002; Holte *et al.*, 2007), while others prefer to classify them by means of embryo grading systems

© 2009 Published by Reproductive Healthcare Ltd, Duck End Farm, Dry Drayton, Cambridge CB23 8DB, UK

(Sharpe-Timms and Zimmer, 2000; Baxter *et al.*, 2006). The above-mentioned differences among laboratories and the existence of diverse classification methods make it difficult to establish assigned values for the embryo images sent to the EQCP for embryo evaluation.

There are many ways in which the assigned values in an EQCP may be established; details are given in ISO 13528:2005 [International Organization for Standardization (ISO), 2005]. For an EQCP in which embryo images are evaluated, there are two possibilities: consensus values from participating laboratories and consensus values from experts.

In other EQCPs in the field of clinical embryology, such as sperm morphology, the assigned values are established from the consensus value from participants (Álvarez *et al.*, 2005), or consensus values from expert laboratories (Björndahl *et al.*, 2002).

In the present paper, the aim was to determine which of the above methods is the most suitable for establishing the assigned value in the EQCP in embryo evaluation.

Materials and methods

All the data utilized in the analysis were obtained from the Spanish EQCP for human reproduction laboratories, organized by Centro de Estudio e Investigación de la Fertilidad (CEIFER, 2009) in association with the ASEBIR. From 2003 to 2007, over 40 laboratories throughout Spain took part in the programme, which was primarily focused on the evaluation of embryo quality and on clinical decision making.

Between 2003 and 2007, a total of 120 embryos were evaluated. Each participating laboratory was sent a DVD/CD-ROM with videos of zygotes and of day 2 and day 3 embryos. Each video was divided into five groups, the first one with five videos of zygotes, the next two groups with five videos each, showing day 2 embryos, and the final two groups with five videos each, showing day 3 embryos. The laboratories were asked to classify each zygote or embryo as optimal, moderate or poor quality. They then had to decide which two zygotes were considered most suitable to retain in culture, and which should be cryopreserved or discarded and, with respect to the embryos, decide for each batch (i.e. day 2 or day 3), which two embryos should be transferred, and of those not transferred, which should be cryopreserved and which should be discarded.

In addition, the ASEBIR Embryo Quality Working Group provided a group of five experts selected by the ASEBIR coordinator. Each of the experts works in a different laboratory, and trained at a different laboratory, centre or university. The experts were asked to evaluate the same videos of embryos used for the Spanish EQCP.

For a given embryo or zygote, the overall classification assigned by the laboratories is that receiving most votes; there was considered to be agreement between the laboratories when, for an embryo, over 75% of the laboratories concurred in their classification or clinical decision. When two embryo classification categories received the same number of votes, the embryo was classed as ‘moderate’. When there were equal opinions on the clinical decision to be taken by laboratories, the embryo was eliminated from the study (this occurred with five embryos).

In the experts’ evaluation, there was considered to be agreement on the classification or clinical decision regarding an embryo or zygote when the five participants made the same choice; if one or more disagreed, there was considered to be disagreement.

For data comparison, the kappa (k) index was calculated to obtain a measure of global agreement, taking into account that which is only to be expected by chance; moreover, this index can be generalized to evaluations of more than two categories. Kappa is intended to give the reader a quantitative measure of the magnitude of agreement between observers (Viera and Garret, 2005). To interpret the level of agreement measured by the kappa coefficient, the proposal made by Landis and Koch (1977) was utilized ($\geq$0.80: almost perfect agreement; 0.61–0.80: substantial agreement; 0.41–0.60: moderate agreement; 0.21–0.40: fair agreement; $\leq$0.20: slight agreement).

Results

The percentage of embryos on which agreement was reached was significantly higher among the experts than among the group of laboratories, both for embryo classification (98.3 versus 44.2%) ($P < 0.001$) and clinical decision (93.3 versus 42.5%) ($P < 0.001$) (**Table 1**).

The agreement between the classification assigned by the laboratories and that determined by the experts presented a kappa coefficient of 0.82 [95% confidence interval (CI): 0.59–1] in the case of zygotes, and of 0.58 (95% CI: 0.39–0.77) and 0.45 (95% CI: 0.24–0.65) for day 2 and day 3 embryos respectively. When the laboratory classification of an embryo was ‘poor’, in no case was the corresponding classification by the experts ‘optimum’; and in only one case (of a day 2 embryo) when the laboratory classification was ‘optimum’ was that of the experts ‘poor’ (**Table 2**). Taking into account just the embryos on which the laboratories were in agreement on the classification, and comparing these results with the experts’ opinions, the corresponding kappa coefficient was 0.70 (0.16–1.00) for zygotes and 0.63 (0.44–0.82) for day 2 and day 3 embryos (**Table 2**).

The agreement between the clinical decision result assigned by the laboratories and that assigned by the experts presented a kappa coefficient of 0.72 (95% CI: 0.43–1) in the case of zygotes, and of 0.57 (95% CI: 0.36–0.78) and 0.46 (95% CI: 0.25–0.67) for day 2 and day 3 embryos respectively. Taking into account just the embryos on which the laboratories were in agreement on the clinical decision, and comparing these results with the experts’ opinions, the corresponding kappa coefficient was 1.00 for zygotes and 0.74 for day 2 and day 3 embryos (**Table 3**).

Table 1. Comparison of the level of agreement concerning embryo evaluation among experts and laboratories ($n = 120$ embryos).

Evaluation parameter	*Experts*	*Laboratories*	P-*value*
Embryo classification	118 (98.3)	53 (44.2)	<0.001
Clinical decision	112 (93.3)	51 (42.5)	<0.001

Values are *n* (%).

Table 2. Comparison of embryo classification (day 2 and day 3 embryos) assigned by experts and laboratories, according to the majority decision of the laboratories or according to only those embryos for which there was high agreement among laboratories.

	Experts' decision		
	Optimal	*Moderate*	*Poor*
Majority decision among laboratories[a]			
Optimal	21 (70.0)	9 (23.7)	1 (3.1)
Moderate	9 (30.0)	24 (63.2)	8 (25.0)
Poor	0 (0.0)	5 (13.2)	23 (71.9)
Total	30	38	32
High agreement among laboratories[b]			
Optimal	12 (75.0)	1 (11.1)	0 (0.0)
Moderate	4 (25.0)	5 (55.6)	3 (14.3)
Poor	0 (0.0)	3 (33.3)	18 (85.7)
Total	16	9	21

Values are *n* (%) or *n*.

[a] Kappa coefficient = 0.52 [95% confidence interval (CI): 0.38–0.65].

[b] Kappa coefficient = 0.63 (95% CI: 0.44–0.82).

Table 3. Comparison of the clinical decision (day 2 and day 3 embryos) assigned by experts and laboratories, according to the majority decision of the laboratories or according to only those embryos for which there was high agreement among laboratories.

	Experts' decision		
	Transfer	*Cryopreservation*	*Rejection*
Majority decision among laboratories[a]			
Transfer	30 (78.9)	9 (27.3)	1 (5.6)
Cryopreservation	4 (10.5)	18 (54.5)	4 (22.2)
Rejection	4 (10.5)	6 (18.2)	13 (72.2)
Total	38	33	18
High agreement among laboratories[b]			
Transfer	17 (89.5)	1 (9.1)	0 (0.0)
Cryopreservation	0 (0.0)	7 (63.6)	1 (9.1)
Rejection	2 (10.5)	3 (27.3)	10 (90.9)
Total	19	11	11

Values are *n* (%) or *n*.

[a] Kappa coefficient = 0.51 [95% confidence interval (CI): 0.36–0.66].

[b] Kappa coefficient = 0.74 (95% CI: 0.56–0.91).

On performing a joint study of the classification and the clinical decision taken for a given embryo, and comparing the results from the laboratories (the category most often assigned) with those from the experts, it can be seen that when most of the laboratories decided that an embryo was of 'poor' quality and not to be transferred, in only 4% (1/25) of cases was cryopreservation recommended (**Table 4**), while for the experts, of the embryos with a 'poor' classification, and hence not suitable for transfer, cryopreservation was recommended in 28% (7/25) of cases ($P = 0.05$) (**Table 4**).

Discussion

On comparing the replies made by the experts and the laboratories, it can be seen that there is almost perfect agreement in the evaluation of zygotes. This finding coincides with the conclusions of Keck *et al.* (2004), who observed that interindividual variability is low concerning the maturation of oocytes and the visualization of pronuclei. Evaluation of day 2 and day 3 embryos revealed a moderate degree of agreement, although this level is low in comparison with the findings of Ziebe *et al.* (2003), who reported a kappa coefficient in the range of 0.82–0.93 among three embryologists concerning evaluation of fragmentation range and cleavage rate. These discrepancies may be partly accounted for by the fact that Ziebe *et al.* (2003) only analysed one characteristic, the fragmentation range, or cleavage rate, while in the present study the whole quality of the embryo was assessed.

In evaluating embryo quality, a three-category classification system (optimal, moderate or poor quality) was adopted, on the basis of its greater simplicity and coincidence with the criteria of Baxter *et al.* (2006). The possible effect on the results of using other embryo classification systems, such as those based on embryo scoring (Desai *et al.*, 2000; Sharpe-Timms and Zimmer, 2000; Fisch *et al.*, 2001; De Placido *et al.*, 2002; Holte *et al.*, 2007) or on embryo grading (Sharpe-Timms and Zimmer, 2000), is not known. Moreover, the comparisons between laboratories are based on the use of video, which means that recording time was limited; in addition, the embryos were not rotated for observation from various angles, and thus the context is an artificial one, in which the embryologist has no control. However, Arce *et al.* (2006) demonstrated the validity of a digital imaging system similar to ours for inte-rembryologist comparisons.

Despite the low kappa coefficients observed, when the experts' opinions did not coincide with those of the laboratory agreements, the difference in classifications was of just one category (i.e. optimal–moderate or moderate–poor), except in one case, in which the experts classified a 2-day embryo as 'poor' and the laboratories as 'optimal'.

Moreover, when the laboratories' response failed to agree with that of the experts' opinion, in the majority of cases (18/32, 56.3%) the quality was overestimated, which is in accordance with the findings of a study concerning the results obtained when sperm morphology is assessed by laboratories lacking the necessary skills for this (Franken and Kruger, 2006). This suggests that laboratories tend to be less rigorous than experts in such evaluations.

Regarding clinical decisions, when the responses made by experts and laboratories were compared, it was found that there was substantial agreement in the evaluation of zygotes, and moderate agreement in the case of day 2 or day 3 embryos. This is in line with Matson (1998), who also found a high degree of variability among embryologists as to the clinical decision to be taken on embryo images. The differences observed arose in a context in which two embryos were transferred; this limitation prevented study of the differences among laboratories with respect to the number of embryos transferred. Matson (1998) observed large differences among the numbers of embryos to be transferred, in a similar multi-centre study,

Table 4. Relationship between embryo classification and clinical decision (day 2 and day 3 embryos) according to the opinion of the experts or according to the opinion of the laboratories.

	Embryo classification		
	Optimal	*Moderate*	*Poor*
Experts			
Clinical decision			
Transfer	24 (80.0)	13 (35.1)	2 (7.4)
Cryopreservation	6 (20.0)	24 (64.9)	7 (25.9)
Rejection	0 (0.0)	0 (0.0)	18 (66.7)
Total	30	37	27
Laboratories			
Clinical decision			
Transfer	23 (85.2)	16 (40.0)	3 (10.7)
Cryopreservation	4 (14.8)	21 (52.5)	1 (3.6)
Rejection	0 (0.0)	3 (7.5)	24 (85.7)
Total	27	40	28

although in this study a very low number of laboratories was used.

The differences observed in the present study concerning the criteria adopted for cryopreserving embryos seem to be responsible for the high degree of variability in the results reported for cryotransfer cycles (Nyboe *et al.*, 2006, 2007, 2008). Clear recommendations should be set out by scientific societies in the field of human reproduction as to which embryos should be cryopreserved and which should not.

There are more discrepancies among the laboratories than among the experts, which is in agreement with the findings of other authors, who have observed that inter-laboratory differences in embryo evaluations are inversely related to the degree of activity, with fewer differences being reported among laboratories with high levels of activity (Baxter *et al.*, 2006), and among experienced embryologists (Arce *et al.*, 2006). The question of intra-observer agreement was not examined; nevertheless, various authors have shown there to be good intra-observer agreement in similar situations (Arce *et al.*, 2006; Baxter *et al.*, 2006). In summary, the low level of agreement between laboratories and experts, as well as the lower variability among the latter, lead to the conclusion that experts' consensus values should be adopted as the assigned values in EQCPs.

Acknowledgements

The authors 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 trust placed in them and the co-operation provided, without which this study would not have been possible.

References

Álvarez C, Castilla JA, Ramírez JP *et al.* 2005 External quality control program for semen analysis: Spanish experience. *Journal of Assisted Reproduction and Genetics* **22**, 379–387.

Arce JC, Ziebe S, Lundin K *et al.* 2006 Interobserver agreement and intraobserver reproducibility of embryo quality assessments. *Human Reproduction* **21**, 2141–2148.

Asociación para el Estudio de la Biología de la Reproducción (ASEBIR) 2008 Criterios de valoración morfológicos de oocitos, embriones tempranos y blastocistos humanos. Cuadernos de Embriología Clínica, Góbalo, Madrid, Spain. p. 59. http://www.asebir.com/publicaciones.htm [accessed 17 September 2009].

Baxter AE, Mayer JF, Shipley SK *et al.* 2006 Interobserver and intraobserver variation in day 3 embryo grading. *Fertility and Sterility* **86**, 1608–1615.

Björndahl L, Barratt CLR, Fraser LR *et al.* 2002 ESHRE basic semen analysis courses 1995–1999: immediate beneficial effects of standardized training. *Human Reproduction* **17**, 1299–1305.

Castilla JA, Ruiz de Assín R, Gonzalvo MC et al. 2009 External quality control for embryology laboratory. *Reproductive BioMedicine Online* **20** [in press].

Centro de Estudio e Investigación de la Fertilidad (CEIFER) Sperm Bank. http://www.ceifer.es/ceifer [accessed 17 September, 2009].

Desai NN, Goldstein J, Rowland DY *et al.* 2000 Morphological evaluation of human embryos and derivation of an embryo quality scoring system specific for day 3 embryos: a preliminary study. *Human Reproduction* **15**, 2190–2196.

De Placido G, Wilding M, Strina I *et al.* 2002 High outcome predictability alter IVF using a combined store for zygote and embryo morphology and growth rate. *Human Reproduction* **17**, 2402–2409.

Fisch JD, Rodriguez H, Ross R *et al.* 2001 The graduated embryo score (GES) predicts blastocyst formation and pregnancy rate from cleavage-stage embryos. *Human Reproduction* **16**, 1970–1975.

Franken DR, Kruger TF 2006 Lessons learned from a sperm morphology quality control programme. *Andrología* **38**, 225–229.

Holte J, Berglund L, Milton K *et al.* 2007 Construction of an evidence-based integrated morphology cleavage embryo score for implantation potential of embryos scored and transferred on day 2 after oocyte retrieval. *Human Reproduction* **22**, 548–557.

ISO 2005 *International Standard ISO 13528. Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparisons*. International Organization for Standardization, Geneva, Switzerland.

Keck C, Fischer R, Baukloh V *et al.* 2004 Quality management in reproductive medicine. In: Gadner DK, Weissman A, Howles CM, Shohan Z, editors *Textbook of Assisted Reproductive Techniques. Laboratory and Clinical Perspectives*. London, New York: Taylor & Francis; 2004. p. 477–494.

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

Magli MC, Van den Abbeel E, Lundin K *et al.* for Committee of the Special Interest Group on Embryology 2008 Revised guidelines for good practice in IVF laboratories. *Human Reproduction* **23**, 1253–1262.

Matson PL 1998 Internal and external quality assurance in the IVF laboratory. *Human Reproduction* **13**, 156–165.

Nyboe A, Goosens V, Ferraretti AP *et al.* 2008 Assisted reproduction technology in Europe, 2004: results generated from European registers by ESHRE. *Human Reproduction* **23**, 756–771.

Nyboe A, Goosens V, Gianaroli L *et al.* 2007 Assisted reproduction technology in Europe, 2003: results generated from European registers by ESHRE. *Human Reproduction* **22**, 1513–1525.

Nyboe A, Gianaroli L, Felberbaum R *et al.* 2006 Assisted reproductive technology in Europe, 2002: results generated from European registers by ESHRE. *Human Reproduction* **21**, 1680–1697.

Practice Committee of the American Society for Reproductive Medicine and the Practice Committee of the Society for Assisted Reproductive Technology 2006 Revised guidelines for human embryology and andrology laboratories. *Fertility and Sterility* **86** (Suppl. 4), 57–72.

Sharpe-Timms KL, Zimmer RL 2000 Oocyte and pre-embryo classification. In: Kal BA, May JV, De Jonge CI, editors *Handbook of the Assisted Reproduction Laboratory*. USA: CRC; 2000. p. 179–196.

Viera AJ, Garret JM 2005 Understanding interobserver agreement: the kappa statistic. *Family Medicine* **37**, 360–363.

Ziebe S, Lundin K, Loft A *et al.* CEMAS II and III Study Group 2003 FISH analysis for chromosomes 13, 16, 18, 21, 22, X and Y in all blastomeres of IVF pre-embryos from 144 randomly selected donated human oocytes and impact on pre-embryo morphology. *Human Reproduction* **18**, 2575–2581.

Declaration: The authors report no financial or commercial conflicts of interest.

Received 5 March 2009; refereed 17 April 2009; accepted 8 September 2009.