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

(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**).