Article

Experimental results on the recognition of embryos in human assisted reproduction

Claudio Manna obtained his MD degree in 1980 and specialization in obstetrics and gynaecology in 1985. He is a University Researcher at the University of Rome 'Tor Vergata', where he shares responsibility for the Infertility Unit. Dr Manna also lectures there on the assisted reproduction technique within the gynaecology specialization course. He is responsible for two centres of assisted reproduction in Rome ('Genesis' and 'Biofertility'). He has produced a multimedia CD-Rom on *Reproduction and Infertility*. He was recently appointed as the Secretary of the Italian branch of the Mediterranean Society for Reproductive Medicine. His particular interest lies in the development of informatics in reproductive medicine.

Dr Claudio Manna

C Manna[1], G Patrizi[2,4], A Rahman[3], H Sallam[3]
[1]Genesis IVF Centre, 00189 Rome, Italy; [2]Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università La Sapienza, 00185 Rome, Italy; [3]Alexandria Fertility Centre, Alexandria, Egypt
[4]Correspondence: Fax: +39 6 4959241; email: patrizi@banach.sta.uniroma1.it

Abstract

The recognition of embryos suitable for transfer in human assisted reproduction is important, and there is evidence that the morphology of the cells may influence the results achievable. A procedure for this recognition problem has been formulated based on morphological attributes of the images of the embryos, and it is therefore useful to compare the recognition of experts with that of a machine programme. The aim of this paper is to compare the precision in the recognition of viable embryos by a group of experts to that of a machine recognition procedure, both for a basic set of embryos and a blind set. Experts were asked to classify the images of 249 embryos transferred to 73 patients, indicated as the training set and another set of 103 embryos transferred to 35 patients called the blind set. A machine programme was used for the same classification. For all the experts the results were statistically not significantly different from independence, which means that viable embryos are not recognized as such for both data sets. Instead, the machine algorithm recognizes in a statistically significant way, the membership class for the embryos submitted. Cell morphology is important for IVF, but differences do not appear to be discernable by the senses, clinical insight, experience and/or training, while classification by machine methods provides more accurate results, which could be improved by enlarging the training set.

Keywords: *automatic selection, embryo selection, human assisted reproduction*

Introduction

IVF is widely used to treat infertility, although the rate of success of treatment is low, with an average overall birth rate per embryo of 13.9% (Templeton *et al*., 1996). Various factors have been identified as affecting this rate, such as the age group of the women concerned, the duration of infertility, the usage of donor's eggs and perhaps the treatment centre and other subsidiary aspects. Other factors have recently been examined, such as the stress suffered by embryonic blastomeres through inappropriate culture conditions (Pool, 2002).

The aim of this paper is to compare the recognition of embryos for their potentiality of producing births as determined by a group of experts and a machine recognition procedure, both for a basic set of embryos, where the main characteristics of the patients are given and a blind set of embryos, where no information is indicated.

There has also been some interest and concern to detect morphological factors of embryonic cells through the application of classification criteria (Puissant *et al*., 1987; Mills, 1992; Steer *et al*., 1992; Giorgetti *et al*., 1995; Saith *et al*., 1998; Racowsky *et al*., 2003). Further, it has been suggested that human embryos can be selected for transfer using morphology at the cleaving and blastocyst stage (Ziebe *et al*., 1997; Scott *et al*., 2000), and many other proposals have been formulated (e.g. Shulman *et al*., 1993) to recognize with precision whether or not the embryo will give rise to a birth (Van Royen *et al*., 1999).

Studies have also been undertaken to determine appropriate scoring methods for day 1, 2, 3 and 5 embryos and combinations of these. In particular, the morphology of pronuclear oocytes has been linked with implantation and development to the blastocyst stage (Scott, 2003).

The selection of suitable embryos may be done through clinical methods (Gardner and Lane, 2003) and expert evaluation (Boiso *et al*., 2002), or through automatic or machine recognition methods, in which a suitable calculus is carried out, either on a-priori clinical principles, or by determining a classifier, i.e. a rule that will consider a set of attributes of an embryo and determine a label for it, which will suggest whether or not it will give rise to a birth (Grimaldi *et al*., 2002; Patrizi *et al*., 2003).

Such a classifier must be regarded purely as an evaluation rule, since no attempt is made to assess the many other aspects that a clinician considers when performing a transfer to a patient. For instance, the evidence for the use in determining quality of oocytes with PB1 and PB2 fluorescence in-situ hybridization is uncertain (Verlinsky *et al*., 2003), but this type of scarcely visible characteristics, or some instrumental or proxy characteristics, may be picked up through the automatic image analysis performed by this algorithm.

The evaluation rule or classifier is formulated on the basis of a set of objects whose class membership is known. On the basis of this set, called a training set, a rule is derived by placing the objects in a homogeneous group on the basis of their attributes and their class. This is an inductive learning procedure, well known to humans, who use it to learn how to read and solve problems and perform diagnoses. The only difference is that this machine algorithm derives explicit mathematical rules, rather than synthetic judgements.

With the automatic recognition procedure, the embryos are classified before transfer, without considering the characteristics of the recipient. However, as it is often indicated, the characteristics of the patients are important aspects for the evaluation of the suitability of the embryo, and thus should also be considered. On the other hand, these aspects are very controversial, so it may be advantageous to inquire regarding the accuracy obtainable by machine classification, without taking these factors explicitly into account.

In fact, what is really being compared is the experts' clinical intuition of the morphology of the embryo and the machine programme's evaluation method, for a reference population. This will not consist of all the potential mothers who would like to be treated, but only of those that are deemed suitable by the clinical centre to receive such transfers and have been accepted for such an intervention.

Materials and methods

Photographs of 249 embryos usually at the 4-cell stage were taken 40–50 h after fertilization and before transfer, by placing each embryo under a microscope (Inverted Microscope Olympus IX 70; Olympus America Inc., Melville, NY, USA), with a camera (video camera JVC TK-C401EG; Victor Company of Japan, Limited (JVC), Yokohama, Japan). Examples of the images used in the experimentation can be checked (http://banach.sta.uniroma1.it/trace/reproduction/initial.html, and follow the links). All the embryos considered were subsequently transferred to uteri.

The research was conducted at a single fertility centre, during the period from January 1998 to October 1999 and involved 73 women, whose infertility factor was mainly andrological, in a total number of 107 transfer cycles with transfer of intracytoplasmic sperm injection (ICSI). The women ranged in age from 23 to 43 years (Grimaldi *et al*., 2002), and later the experiment was extended to over 800 patients (Patrizi *et al*., 2003).

Ovarian stimulation was carried out by administering recombinant FSH (Gonal F; Serono International SA, Geneva, Switzerland) at a dosage of 150–400 IU according to individual response after suppression with gonadotrophin-releasing hormone analogue in a daily preparation (Suprefact; Hoechst Marion Roussel Deutschland, Bad Soden Germany). Oocyte retrieval was carried out with ultrasound-guided transvaginal follicular aspiration after 35 h from the administration of 5000 IU of human chorionic gonadotrophins (Profasi, Serono). Gametes and embryos were cultured under oil in drops of a culture medium (IVF Scandinavia; Vitrolife Sweden AB, Kungsbacka, Sweden) with an atmosphere of CO_2 of 5% in air. The ICSI process was performed according to current methodology (Van Steirteghem *et al*., 1993).

The control (blind) set consisted of 35 women whose characteristics were not significantly different from the initial sample.

With the training set of embryos, the number of births obtained was 32 from 27 women, while for the blind set 15 births occurred from 13 women. For the two samples, the average number of births per embryo transferred was 12.85 and 14.6% respectively. The two proportions are not significantly different from each other or from overall birth rates, as indicated above, at a 95% statistical significance level and are in line with common experience.

Six experts in the field of embryo selection and transfer, were chosen and were asked to consult the given web page, or a CD-ROM with the two data sets, which had been forwarded to them. The two data sets consisted of images of embryos forming the training set and the blind set. The first set contained some information on the prospective mother, such as age, while the second set did not include any information. Two sets of tests were run. The first test asked the experts to indicate the number of births per transfer cycle from the images of the given embryos, irrespective of which embryo gave rise to a birth, while the second test asked to indicate the actual embryos that could have given rise to the births that had been predicted in the first test.

More accurate predictive results can be obtained by using more clinical centres, randomly drawn, and a larger random sample, but the experimental design adopted, here, in which only one clinical centre was used and the sample consisted of all the transferred specimens in the period, may be considered adequate for this experimentation.

It was pointed out to the experts that the data submitted consisted of an experimental sample, and might not have come from a typical population, due to the location of the centre, the historical period of time and other factors, which

could affect the sample drawn. Also, the double task assigned to the experts blocked the rather obvious policy to assign no births, on the presumption that the birth rate would be much lower (it was in fact around 14%), so with such a policy, the precision of a correct recognition increases to 86%.

In fact, no expert fell back to this stratagem, although one expert (F) did predict that each patient would have at least one successful birth for one data set, but not for the other. Thus even in this case it can be assumed that his choices reflected his evaluations, as otherwise he would have used the same policy for both sets.

Every expert was asked to record in suitable electronic forms, the following assessments: (i) for every transfer cycle indicate how many births have occurred from these embryos. Thus, a number 0, 1, 2,... had to be written next to each identifier of each patient and each cycle. This was done for each entry in the training set and in the blind set; (ii) for every transfer cycle, indicate the actual embryos that gave rise to a birth. Thus if a patient supposedly was going to give rise to two births, the request was to choose which two embryos would be responsible for this occurrence. Again, this was done for the two sets of data.

While the outcome of the first test is known with certainty (the number of births occurring in each transfer cycle), the imputation of which embryo was in fact responsible for that birth is not certain, but can be conjectured on the basis of the automatic algorithm, which indicates this. Thus the second test must be regarded as a consistency check on the machine selection process.

Their answers were collected and the contingency tables were formed; see **Tables 1** and **2** for the first task and **Tables 3** and **4** for the second task.

Table 1 for the training set and **Table 2** for the blind set were compiled as follows: 1. For every transfer cycle and for each expert, the number of births was predicted by the expert and this was compared with the number of actual births. 2. If one birth had occurred and one baby had been declared, a '1' was scored for that expert and that transfer cycle in the cell of the computational table in which one box was identified by the labels 'birth' and 'birth predicted'. 3. If twin births had occurred and only one had been predicted, a '1' was scored as before while in the box labelled by 'birth' and 'no-birth predicted' a second '1' was scored. 4. If no births had occurred but one was predicted for that transfer cycle, then a '1' was scored in the box 'no-birth'; and 'birth predicted';. 5. Similarly, if no births were predicted and no births occurred, for that set of embryos, the number in that set was scored. 6. The scores obtained for each expert were then summed and transferred to a contingency table with four entries as depicted in **Tables 1** and **2**. It should be noted that the margin totals varied because of the different imputations made by the experts.

For **Tables 3** and **4**, exactly the same procedure was followed, this time on the individual embryos. The total number of embryos were respectively 249 for the training set and 103 for the blind set, so these numbers formed the marginal grand total, as these are the number of embryos considered in each set, however they were assigned.

The predicted outcome for each embryo by each expert, which formed the basis of the scoring in these tables, was compared with those given by TRACE (total recognition by adaptive classification experiments) algorithm.

The automatic recognition algorithm uses the training set to form rules to assign embryos to either class based on their image. The working of the algorithm is intuitive, and is based on the human induction process, which it carefully follows. The images considered are transformed into a set of attributes, or characteristics, on the basis of their pixel map and to each image in the training set a label is attached. The algorithm then will form as few homogeneous groups as possible so that all the members of each group have the same label. The mean set of attributes for each group is then used to assign labels to unlabelled objects. Each unlabelled object will receive the label of the group whose mean set of attributes is closest to its set of attributes. See the Appendix for all the details.

All that is important here is to grasp that there is a machine that recognizes with a certain performance and can be compared with the live experts. The results of the machine algorithm are presented in **Table 5** for the first test for both the training set and the blind set and in **Table 6** for the second test.

Results

Fisher's exact test (Conover, 1971) and the log odds ratio test were carried out for each expert's 2 × 2 contingency tables presented in **Tables 1–4** and for the machine programme results presented in **Tables 5** and **6**. Additionally, the log odds ratio test was also applied (Agresti, 1990) and all the results of these statistical tests are presented in **Table 7** and **Table 8**.

Fisher's exact test works as follows. The frequencies in the table may have arisen independently or according to a systematic evaluation regarding the suitability of embryos, so the entries will not be independent. If the marginal row and column totals are fixed for a given table, the different possible tables that can occur can be characterized by the frequency that will arise in the top left cell. It is therefore possible to determine for all possible values of this frequency, what would be its probability that such a value had arisen, if there was independence. Thus if the probability of the outcome frequency is higher than 0.05, the entry of the table will be considered independent. Instead, if the probability is lower than that level, this hypothesis of independence must be rejected in favour of a systematic relationship (Agresti, 1990).

The log odds ratio test is similar, but it is an asymptotic one. It is based on comparing the odds of probabilities of the entries that have occurred and one resorts to natural logarithms, so that the product ratio is quickly distributed like a normal distribution in the case of independence with a zero mean and an easily calculated standard error. Alternatively, if the assignments of the embryos to the class had been systematic, the mean of the log odds ratio would mark a significant departure from a zero value. When normalized by the variance, the normalized log odds ratio will be either higher than 1.96, or smaller than –1.96.

Table 1. Classification of births by transfer cycles in the training set: results by experts.

	Birth	*No birth*	*Total*
Predictions of A			
Birth	19	13	32
No birth	27	28	55
Total	46	41	87
Predictions of B			
Birth	10	22	32
No birth	23	33	56
Total	33	55	88
Predictions of C			
Birth	13	19	32
No birth	15	36	51
Total	28	55	83
Predictions of D			
Birth	16	16	32
No birth	25	30	55
Total	41	46	87
Predictions of E			
Birth	20	12	32
No birth	23	34	57
Total	43	46	89
Predictions of F			
Birth	30	2	32
No birth	81	0	81
Total	111	2	113

Table 2. Classification of births by transfer cycles in the blind set: results by experts.

	Birth	*No birth*	*Total*
Predictions of A			
Birth	7	8	15
No birth	12	13	25
Total	19	21	40
Predictions of B			
Birth	2	13	15
No birth	9	15	24
Total	11	28	39
Predictions of C			
Birth	3	12	15
No birth	8	16	24
Total	11	28	39
Predictions of D			
Birth	5	10	15
No birth	12	13	25
Total	17	23	40
Predictions of E			
Birth	7	8	15
No birth	10	15	25
Total	17	23	40
Predictions of F			
Birth	10	5	15
No birth	25	8	33
Total	35	13	48

Thus, if the log odds ratio lies near zero, the frequencies are said to be independent, as before by performing the required test of significance.

These tests of hypothesis compare the actual outcome with a theoretical outcome arising from independence of the entries and may be subject to some random error. These tests are therefore devised to ensure that if the data are really independent, then in only, say, 5% of cases will the test reject this conclusion. Obviously a type II error can also occur, and the entries would be considered independent when really they are not.

Note in any one trial, the null hypothesis is either correct or wrong, but this cannot be determined. What is known is that if the test is done over and over again with similar data, in 95% of the times that the null hypothesis is confirmed, it will be correct.

The attribution of the experts as to the number of births that would result from the given transfers was independent of the characteristics of the embryos. Thus, embryos were classed with the same probability as, viable or not, irrespective of their actual characteristics.

If there is no recognition capability on the part of the expert, the number of times an embryo will be assigned to the 'birth predicted class' or to the 'no birth predicted' class is about the same, and this can be indicated as statistically independent. Instead, if the expert is able to determine the potentiality of the embryo correctly, the frequencies will be mainly concentrated along the main diagonal of the 2 × 2 contingency table constructed for each expert, as indicated above. This is not the case, as can be seen from tables.

The same conclusions are reached for the expert selections in the blind set in **Table 2**. Thus, the tests indicate that for all the experts, there is no evidence that the recognition is systematic.

Much the same results occur in **Table 3** and **4** for all the experts, except expert E. This case is important and will be examined below.

An interesting point is raised by expert F, who predicted at least one birth from every transfer cycle in the training set. This could have been considered as a misunderstanding of the instructions regarding the selection process and scoring methods, but this is unlikely, since the data given for the blind set reflects a correct interpretation of the instructions.

Table 3. Classification of embryos by outcome in the training set: results by experts.

	Birth	*No birth*	*Total*
Predictions of A			
Birth	9	23	32
No birth	37	180	217
Total	46	203	249
Predictions of B			
Birth	4	28	32
No birth	29	188	217
Total	33	216	249
Predictions of C			
Birth	5	27	32
No birth	23	194	217
Total	28	221	249
Predictions of D			
Birth	5	27	32
No birth	36	181	217
Total	41	208	249
Predictions of E			
Birth	10	22	32
No birth	33	184	217
Total	43	206	249
Predictions of F			
Birth	18	14	32
No birth	93	124	217
Total	111	138	249

Table 4. Classification of embryos by outcome in the blind set: results by experts.

	Birth	*No birth*	*Total*
Predictions of A			
Birth	5	10	15
No birth	14	74	88
Total	19	84	103
Predictions of B			
Birth	1	14	15
No birth	10	78	88
Total	11	92	103
Predictions of C			
Birth	3	12	15
No birth	8	80	88
Total	11	92	103
Predictions of D			
Birth	5	10	15
No birth	12	76	88
Total	17	86	103
Predictions of E			
Birth	6	9	15
No birth	11	77	88
Total	17	86	103
Predictions of F			
Birth	8	7	15
No birth	27	61	88
Total	35	68	103

Table 5. Classification of births by transfer cycles in both sets: results by the automatic machine programme.

	Training set predictions			*Blind set predictions*		
	Birth	*No birth*	*Total*	*Birth*	*No birth*	*Total*
Birth	25	7	32	4	11	15
No birth	26	30	56	3	20	31
Total	51	37	88	7	31	38

Table 6. Classification of embryos in each set: results by the automatic machine programme.

	Training set predictions			*Blind set predictions*		
	Birth	*No birth*	*Total*	*Birth*	*No birth*	*Total*
Birth	16	16	32	4	11	15
No birth	35	182	217	3	85	88
Total	51	198	249	7	96	103

Table 7. Probability values for Fisher's exact test for the classification of outcomes from the training and blind sets for transfer cycles and single embryos.

Expert	*Outcome by transfer cycle*		*Outcomes by embryos*	
	Training set	*Blind set*	*Training set*	*Blind set*
A	0.3817	1.0000	0.1451	0.1457
B	0.4927	0.1504	1.0000	1.0000
C	0.3441	0.4770	0.3759	0.1993
D	0.8241	0.5115	1.0000	0.1238
E	0.0505	0.7486	0.0414	0.1068
F	0.0784	0.5088	0.1837	0.1378
Machine	0.0066	0.4008	5.05×10^{-5}	0.0082

Table 8. Log odds ratio results and their standard error of estimates for the classification of outcomes fro the training and blind sets for transfer cycles and single embryos.

Expert	*Training set for transfer cycles classification*		*Blind set for transfer cycles classification*		*Training set for embryo classification*		*Blind set for embryo classification*	
	Log odds ratio	*Standard error*	*Log odds ratio*	*Standard error*	*Log odds ratio*	*Standard error*	*Log odds ratio*	*Standard error*
A	0.4159	0.4498	–0.0535	0.6543	0.0181	0.4326	0.9719	0.6204
B	–0.4274	0.4683	–1.3610	0.8687	–0.0768	0.5705	–0.5849	1.0882
C	0.4960	0.4836	–0.6931	0.7773	0.4460	0.5345	0.9163	0.7444
D	0.4823	0.4453	–0.6131	0.6784	–0.0714	0.5199	1.1527	0.6297
E	0.9017	0.4541	0.2719	0.5592	0.9200	0.4257	1.5404	0.6178
F	–	–	–0.4463	0.6819	0.5390	0.3818	0.9496	0.5668
Machine	1.4161	0.5046	0.8855	0.8510	1.6487	0.3988	2.3324	0.8283

If an expert classifies all the patients as not having any successful births, then he will be correct in a very large number of cases, but he will destroy his livelihood. For the same reason, the machine programme must not apply these basic statistical outcomes, since in this case it will report no births. The importance of the two tests performed is now evident.

A similar analysis is conducted on the classification of individual embryos indicated in **Tables 3** and **4**.

The classification results for the algorithm TRACE are reported in **Tables 5** and **6** with regard to the predicted outcomes per transfer cycle and with regard to the prediction of which embryos will gave rise to births.

For the prediction of the elements in the training set, replication of the training was carried out 150 times by first randomly selecting 10% of the training set and placing it aside for verification and then performing the training until convergence, in which all the patterns were classified correctly. It is important to note that training is always achieved with complete accuracy, that is all the patterns in the training set are assigned by the algorithm to the same class as they were assigned originally by the 'teacher' or by their outcome (here birth or no birth; Nieddu and Patrizi, 2000).

Having trained the algorithm, it was applied to the randomly selected verification set, without considering the known class label and the algorithm attributed a label, based on a least distance criterion as described in the Appendix. The result was compared with the known class label and was used as the basis for the results reported in the tables.

As this verification procedure was repeated 150 times, on average every embryo appeared in verification 15 times, so each was placed in class 1 or 2, depending on the majority of times that it was assigned to one of the two classes (birth, no birth).

It may happen that some embryos, due to chance, did not appear exactly 15 times, but an even number of times and the image was classified in verification in each class an equal number of times (ties). Thus, a residual verification run was carried out, with just those images of embryos that had resulted in a tie in the classification forming the verification set. In this way, the recognition of every embryo in the training set was also determined by the algorithm, and these are the results reported in aggregate form for the patients in **Table 5** and for each embryo in **Table 6**. The Fisher Exact test carried out on this data is highly significant in most cases. In fact, for the training set, the probability of a two-sided independence was less than 0.0066 and this confirms that there is dependence between the results reported.

In the identification of each embryo as being viable or not, the results are significant in the training set, which exhibits a very low probability that this result could have occurred by chance (5.05×10^{-5}) or less than one chance in 20,000.

For the blind set, the whole training set was used to train the classifier, and the classifier that resulted was used to classify the patterns in the blind set consisting of the images of 103 embryos. The results are reported in **Tables 5** and **6** as well respectively for the patients and for the embryos transferred.

In **Table 5** for the blind set, Fisher's exact test revealed that the results were not significantly different from independence, while for **Table 6**, where recognition was carried out on each embryo, the Fisher Exact test provided a very significant result indicating a value of 0.0082. The classification procedure being different for the training set and the blind set may lead to some differences over and above sampling variability, due to this factor. In particular, the low frequencies obtained for the blind set in **Table 5** may account for this non-significant result.

This aspect is intrinsic in the statistical tests and is comforting, as it shows that these tests are very sensitive.

However, in this case, if a bigger set had been examined all these results might have been significant. It is well known that cell frequencies should be greater or equal to 5 (Haberman, 1978). This was in fact brought out in the extension of the research conducted (see Patrizi *et al*., 2003).

Discussion

These experiments were carried out with a small sample, and more accurate results could be obtained by extending the sample of embryos to be used in the experiment (see Nieddu and Patrizi, 2000 and Patrizi *et al*., 2003). Thus, the prediction error can be made increasingly small by suitably enlarging the training sample. This will render false negative results increasingly unlikely, although this is not an issue in this paper.

The indication by all the experts of the number of births that will arise in a given circumstance from the individual transfer cycles, on the basis of the evidence given here, is no better than if a random draw, while five out of six experts failed in the selection of embryos that are considered responsible for a birth. In contrast, for the algorithm TRACE, precision is significantly different from randomness at the given level of confidence, except for the recognition of the number of births by transfer cycle in the blind set, which is probably due to the frequencies resulting after the aggregation by transfer cycle, technically because of the reduction in the degrees of freedom.

Expert E behaves differently. With regard to the determination of the number of births for the patients in the training set, his classification is not significantly different to that obtainable by random choice, while in the identification of the embryos, the accuracy is significantly different from equiprobality.

Different selection criteria seem to inspire expert E in his choice. Perhaps he relies too much on prior knowledge rather than observation. In fact, expert E has a close association with the experimentation, and the fact that the analysis undertaken provides different results for this expert is an important indication of the accuracy and the sensitivity of the analysis. It is for this reason that his results have been included.

Clearly, the behaviour of the six experts in the various tests indicates that selection is a value laden choice process, or ideological and that not enough weight is given to the systematic factors, whatever they are, which permit recognition. There is emotion in their choice, which is not present with TRACE. This favours machine methods for diagnosis, perhaps in combination with clinical insight.

Moreover, the machine procedure does not require visualization, in the same way as humans, of the morphological factors to be considered, and it could well be that, for instance, when considering bovine embryos, the determination of the attributes would not be jeopardized by the opacity of the blastomeres (Van Soom *et al*., 2003).

To conclude, the statistical tests performed indicate the superiority of the algorithm over the judgement of experts. Larger samples will give better results, but the general conclusion should be nonetheless clear.

References

Agresti A 1990 *Categorical Data Analysis*. Wiley, New York City, USA.

Boiso I, Veiga A, Edwards RG 2002 Fundamentals of human embryonic growth in vitro and the selection of high-quality embryos for transfer. *Reproductive BioMedicine Online* **5**, 328–350.

Bonifazi G, Massacci P, Nieddu L, Patrizi G 1996 The classification of industrial sand-ores by image recognition methods. In: *Proceedings of the 13th International Conference on Pattern Recognition Systems*, *volume 4: Parallel and Connectionist Systems*. Computer Society Press, Los Alamitos, CA, USA, pp. 174–179.

Conover WJ 1971 *Practical Nonparametric Statistics,* Wiley, New York.

Gardner D, Lane M 2003 Towards a single embryo transfer. *Reproductive BioMedicine Online* **6**, 470–481.

Giorgetti C, Terriou P, Auquier P *et al*. 1995 Embryo score to predict implantation after in-vitro fertilization: based on 957 single embryo transfers. *Human Reproduction* **10**, 2427–2431.

Grimaldi G, Manna C, Nieddu L *et al*. 2002 A diagnostic decision support system and its application to the choice of suitable embryos in human assisted reproduction. *Central European Journal of Operational Research* **10**, 29–44.

Haberman S 1978 *Analysis of Qualitative Data* (2 volumes). Academic Press, New York.

Mills CL 1992 Factors affecting embryological parameters and embryological selection for IVF–ET. In: Group TPP (ed.) *In Vitro Fertilization and Assisted Reproduction*. Parthenon, London, pp. 187–204.

Nieddu L, Patrizi G 2000 Formal properties of pattern recognition algorithms, a review. *European Journal of Operational Research* **120**, 459–495.

Patrizi G, Manna C, Moscatelli C, Nieddu L 2003 Pattern recognition methods in human-assisted reproduction. *International Transactions in Operational Research* **10**, 1–15.

Pool T 2002 Recent advances in the production of viable human embryos in vitro. *Reproductive BioMedicine Online* **4**, 294–302.

Puissant F, Van Rysselberge M, Barlow P *et al*. 1987 Embryo scoring as a prognostic tool in IVF treatment. *Human Reproduction* **3**, 705–708.

Racowsky C, Combelles C, Nureddin A *et al*. 2003 Day 3 and day 5 morphological predictions of embryo viability. *Reproductive BioMedicine Online* **6**, 323–331.

Saith RR, Srivivasan A, Mitchie D, Sargent IL 1998 Relationships between the developmental potential of human IVF embryos and features describing the embryo, oocyte and follicle. *Human Reproduction Update* **4**, 121–134.

Scott L 2003 Pronuclear scoring as a predictor of embryo development, *Reproductive BioMedicine Online* **6**, 201–214.
Scott L, Alvero R, Leondires M, Miller B 2000 The morphology of human pronuclear embryos is positively related to blastocyst development and implementation. *Human Reproduction* **15**, 2394–2403.
Shulman A, Ben-Nun I, Ghetler Y *et al*. 1993 Relationship between embryo morphology and implantation rate after in-vitro fertilization treatment in conception cycles. *Fertility and Sterility* **60**, 123–126.
Steer CV, Mills CL, Tan SL *et al*. 1992, 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. *Human Reproduction* **7**, 117–119.
Templeton A, Morris JK, Parslow W 1996, Factors that affect outcome of in-vitro fertilisation treatment. *Lancet* **348**, 1402–1406.
Van Royen E, Mangelschots K, De Neubourg D *et al*. 1999 Characterization of top quality embryo, a step towards a single-embryo transfer, *Human Reproduction* **14**, 2345–2349.
Van Soom A, Mateusen B, Leroy J, de Kruiff A 2003, Assessment of mammalian embryo quality: what can we learn from embryo morphology. *Reproductive BioMedicine Online* **7**, 664–670.
Van Steirteghem AC, Nagy Z, Joris H *et al*. 1993, High fertilization and implantation rates after intracytoplasmic sperm injection. *Human Assisted Reproduction* **8**, 1061–1066.
Verlinsky Y, Lerner S, Illkevitch N *et al*. 2003 Is there any predictive value of first polar body morphology for embryo genotype of development potential? *Reproductive BioMedicine Online* **7**, 336–341.
Watanabe S 1985 *Pattern Recognition: Human and Mechanical*. Wiley, New York City, USA.
Ziebe S, Petersen K, Lindenberg S *et al*. 1997, Embryo morphology or cleavage stage: how to select the best embryos for transfer after in-vitro fertilization, *Human Reproduction* **9**, 1545–1549.

Appendix

The classification algorithm consists of a procedure called TRACE (Total Recognition by Adaptive Classification Experiments; Nieddu and Patrizi, 2000). Here embryos are identified as belonging to one type or to the other, as in the earlier study, where the algorithm and the results are fully explained (Grimaldi *et al*., 2002), while a larger experiment was also conducted (Patrizi *et al*., 2003).

The pattern recognition algorithm

Pattern recognition methods consider a number of defined characteristics, measured for each object to form a pattern vector, which is an ordered list of values of the chosen characteristics for that object. Each object (embryo) considered is assigned to a membership class, available after exhaustive tests or at the end of a certain period (e.g. gestation). The class to which the object belongs may eventually be known precisely or imprecisely, i.e. with a certain ambiguity.

The aim of a pattern recognition algorithm is to determine a rule, called a classifier, such that given the characteristics of an object, it will assign to it a membership class, on the basis of the rule. Thus, it constitutes an induction procedure, where the membership class is inferred, based on the available characteristics (Nieddu and Patrizi, 2000).

The method used, depicted by a flow diagram in **Figure 1** considers a training set of patterns with assigned membership classes to form the classifier. For each class, a mean pattern vector, called a barycentre, is formed component by component, by averaging over each characteristic for all the patterns in that class.

Once these barycentres have been calculated, one for each class, the distance of every pattern from each barycentre is calculated. All patterns which fall nearer to a barycentre of another class than to a barycentre of its own class are marked. The pattern that is the furthest from the barycentre of its own class is selected and used as a seed for a new sub-group of the class.

Distances of all the patterns from all the barycentres of the given class and the new seed barycentre are determined and each pattern is reassigned to the barycentre of its own class to which it is nearest. The barycentres are calculated anew with just the patterns assigned to that barycentre, which are consequently updated.

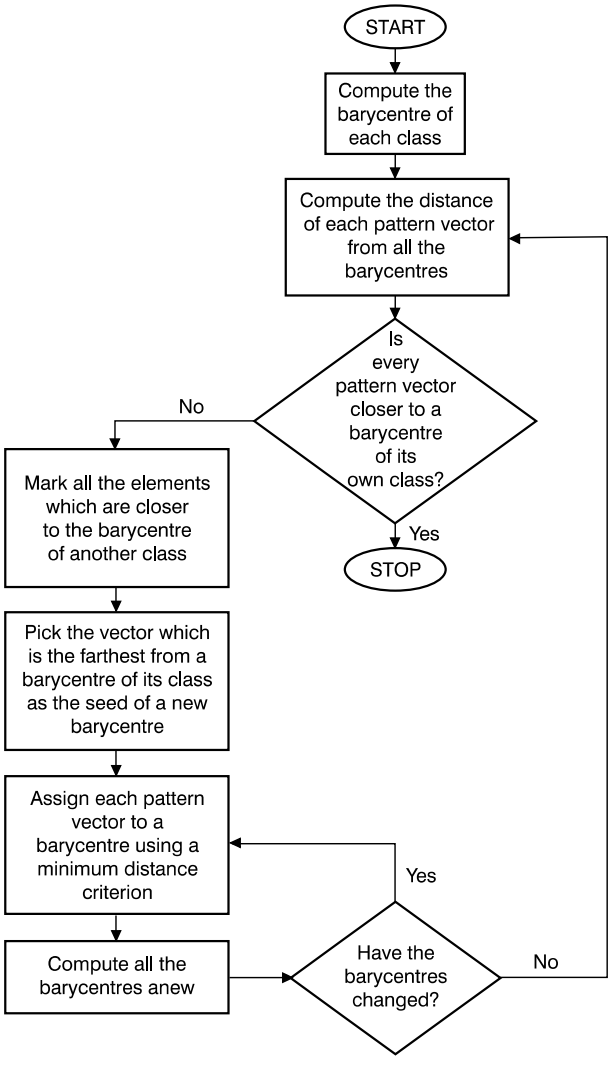

Figure 1. Flow-chart of TRACE algorithm.

The procedure continues, with one additional barycentre. Again patterns are marked, as above, a new seed barycentre is determined and the calculations are repeated with a higher level of homogeneity.

It is easy to show that if there are no two identical patterns assigned to different classes, a homogeneous set of barycentres is formed, such that each pattern is closer to a barycentre from its own class than to one of the other class (Nieddu and Patrizi, 2000). When this is achieved, the classifier has been determined.

In classification, the set of barycentres is used to determine which barycentre is nearest to the pattern to be classified, and then the class of that barycentre is assigned to the pattern.

In a verification phase, part of the training set is set aside, usually by selecting randomly, say 10% of the training set, to use as a classification set, by ignoring the label of each object, and the classifier is formed from the other 90% of the training set. Then the membership classes for the objects in the verification set are determined, as if the verification set constituted a set to be classified and the class assigned to each pattern is then checked with their real class attribution, to determine the precision of the algorithm.

The number of trials and the training set are conventional, but it is desired to choose sizes that allow a small standard error without reducing the training set too much.

Determination of the properties of the data set

In this experiment regarding embryos, the characteristics of the objects are defined from the images of the embryos, by considering each image as an array of intensities of shades of grey, from which the frequency distribution of the grey tones for each picture in the horizontal and vertical direction can be calculated. This pixel-intensity profile indicates the homogeneity of the image, or whether it has many dark and light spots and how they are distributed. From these distributions, a certain number of polynomial functions of central tendency and spread can be calculated, called central moments, to express the shape of the distribution in a standardized way. Here the first five moments, neglecting the first, which is identically zero, in the vertical and horizontal direction are considered for a total of 10 attributes.

An image was assigned to class 2 if a birth occurred from that embryo, and to class 1 otherwise. Classes could initially only be assigned with certainty in the case of single embryo transfers, or in those cases where all the transferred embryos gave rise to no births or when a set of embryos gave rise to the same number of births.

Multiple transfers are the most common occurrence in which the number of embryos is greater than the number of births, so this must be resolved by determining which embryos are likely to be responsible for the births. This was done in the following way.

Apply this algorithm to the embryos whose outcomes are known with certainty. These sure instances are used as the training set and all the other pattern vectors were considered as yet not classified, so after having formed the classifier, a classification can be undertaken on this set of patterns to be classified.

If for any group of embryos transferred to the same uterus, the number of patterns classified as belonging to class 2 is greater than the declared number of births, one or more patterns which were furthest from the closest barycentre of class 2 are arbitrarily assigned to the other class, and *vice versa*. Through the application of this algorithm, every embryo is assigned to a specific class. This procedure was performed only once.

The final result of this stage is that every embryo has received a unique class label, although the classification must be considered as achieved with an 'imprecise' teacher (Watanabe, 1985).

The correction algorithm

As some embryos may have been incorrectly classified due to the rather arbitrary way in which multiple transfers are assigned, a second stage of recognition is executed, by using a suitable correction process.

Consider an object that has been placed in the wrong class by the 'teacher' and suppose that it is used in the verification set. The object will be assigned predominantly to its correct class, if the classification is assumed to have a good precision, as its class label is not considered and as only its attributes are considered for the classification. However, since by hypothesis it was assigned initially to the wrong class, in the verification, on repeating this experiment repeatedly with different training sets, it will result predominantly assigned by the algorithm to a wrong class.

By replicating the training and verification phase, say 150 times, as indicated above, 150 replications are obtained and on the average a given pattern will appear in verification about 15 times. This is quite sufficient to base a conservative correction procedure.

Thus, it is methodologically sound to correct the classification, if the misclassification occurs in a high proportion of the time that the object appears in the verification sample, significantly more often than it would be misclassified randomly by the algorithm. It is considered correct to change the class if a pattern is misclassified in verification, over 66% of the time, given that the average rate of precision here is over 78%, see (Bonfazi *et al*., 1996).

Here it is required, since the number of births is fixed for each patient, to impose an additional condition. A corresponding embryo transferred to the same uterus from another class had to be similarly misclassified. Thus if for a transfer cycle it was found that in a high proportion of the times two embryos were classified incorrectly in verification in different classes, the assigned classes were exchanged for the two embryos. This was carried out on eight pairs of embryos regarding seven women.

Full details regarding the classification algorithm and the imprecise training and correction phase have been published elsewhere (Grimaldi *et al*., 2002; Patrizi *et al*., 2003).

Received 24 October, 2003; refereed 7 November, 2003; accepted 8 January, 2004.