Is there an advantage in scoring early embryos on more than one day?

Catherine Racowsky¹, Lucila Ohno-Machado²,³, Jihoon Kim², and John D. Biggers⁴

¹Department of Obstetrics, Gynecology and Reproductive Biology, Brigham and Women’s Hospital, Harvard Medical School, 75 Francis Street, ASB 1+3, Rm 082, Boston, MA 02115, USA ²Decision Systems Group, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA 02115, USA ³Division of Biomedical Informatics, Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA ⁴Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA

Correspondence address. Tel: +1-617-732-5455; Fax: +1-617-975-0825; E-mail: cracowsky@partners.org

BACKGROUND: This study was undertaken to determine what characteristics should be recorded on which days to build a predictive model for selection of Day 3 embryos.

METHODS: Embryos failing to form a clinical sac or that formed a viable fetus (to ≥12 weeks), and transferred singly (n = 269) or in pairs (n = 1326) were scored for early cleavage and pronuclear status on Day 1, and cell number, fragmentation, and symmetry on Days 2 and 3, with number of nuclei per blastomere also recorded on Day 2. Seven candidate models were identified using a priori clinical knowledge and univariate analyses. Each model was fit on a training-set and evaluated on a test-set with resampling, with discrimination assessed using the area under the ROC curve (AUC) and calibration assessed using the Hosmer–Lemeshow statistics.

RESULTS: Models built using Day 1, 2 or 3 scores independently on the 30 resampled data sets showed that Day 1 evaluations provided the poorest predictive value (median AUC = 0.683 versus 0.729 and 0.725, for Day 2 and 3). Combining information from Day 1, 2 and 3 marginally improved discrimination (median AUC = 0.737). Using the final Day 3 model fitted on the whole dataset, the median AUC was 0.732 (95% CI, 0.700–0.764), and 68.6% of embryos would be correctly classified with a cutoff probability equal to 0.3.

CONCLUSIONS: Day 2 or Day 3 evaluations alone are sufficient for morphological selection of cleavage stage embryos. The derived regression coefficients can be used prospectively in an algorithm to rank embryos for selection.

Key words: embryo morphology / embryo score / evaluation day / implantation potential / prediction

Introduction

A great challenge in clinical IVF is to refine embryo selection techniques so that the single most developmentally competent embryo can be reliably identified in every cohort of available embryos. Once this challenge has been met, we may be positioned to perform single embryo transfer to all patients. As such, IVF pregnancy rates may be maximized, and IVF-related multiple pregnancies may be eliminated except for those that very rarely occur due to embryo splitting.

Historically, morphological evaluation has been the primary method used for embryo assessment (Edwards et al., 1981; Cummins et al., 1986) and, despite its recognized limitations (Guerif et al., 2007), the method still remains the most commonly used approach for selection. It has been recognized for many years that a large proportion of human preimplantation embryos undergo deviant development in vitro, failing to follow the expected normal developmental timeline (Puissant et al., 1987; Staessen et al., 1992; Steer et al., 1992; Van Royen et al., 1999), and frequently exhibiting high levels of fragmentation and blastomere asymmetry. These facts have provided the rationale for investigating whether multiple evaluations through early preimplantation development may improve selection compared with a single evaluation, performed shortly before transfer. A large number of studies have been performed in which various combinations of days for scoring have been assessed (Skiadas and Racowsky, 2007 for review) and systems for evaluation have been proposed (e.g. Cummins et al., 1986; Puissant et al., 1987; Steer et al., 1992; Giorgi et al., 1995; Van Royen et al., 1999; Desai et al., 2000).

Regardless of the days on which embryos have been assessed, several approaches have been made to establish various numerical scoring systems to predict the likelihood of an embryo giving rise to a viable fetus. The methods can be classified as follows: morphological observations and the allotment of scores (Giorgi et al., 1995; Desai et al., 2000; Van Royen et al., 1999, 2001; Fisch et al., 2003; Sjöblom et al., 2006); the application of logistic regression analysis (Terriou...
et al., 2001, 2007; Guerif et al., 2007; Holte et al., 2007; Vergouw et al., 2008); class probability tree analysis (Saith et al., 1998); a case based reasoning system (Jurisica et al., 1998); decision tree data mining (Trimarchi et al., 2003) and automated pattern analysis (Patrizi et al., 2004). Overall, conflicting results exist and there is currently no consensus on (1) the optimum day(s) for evaluation; (2) the optimum set of variables that should be used in a predictive model; and (3) model performance benchmarks for future comparison. Answers to these questions are necessary so that quantitative comparisons can be made about the relative efficacies of the different evaluation protocols.

The present study was undertaken to identify the utility of embryo scoring on Day 1 for early cleavage on the afternoon of the fertilization check, and on Day 2 and Day 3. The over-arching goal of the research was to determine what days the embryos should be scored and what characteristics should be recorded, so as to achieve a ranking of available embryos for purposes of selection for transfer.

Materials and Methods

The study was approved by the Partners’ Institutional Review Board for medical record review of our electronic IVF database.

Ovarian stimulation and luteal support protocols

Patients having IVF, with or without ICSI, underwent controlled ovarian stimulation (COS) using protocols previously described in Skiadas et al. (2006). Briefly, COS was most commonly performed using luteal leuprolide acetate (Lupron; TAP Pharmaceuticals, Deerfield, IL, USA) down-regulation in conjunction with either highly purified FSH (Fertinex; Serono Laboratorones, Norwell, MA, USA) or recombinant FSH (Follistim: Organon, West Orange, NJ, USA; Gonal-F: Serono Laboratories). The standard daily gonadotrophin dosage was typically three to four ampules (225–300 IU) administered as either a single or split dose. However, patients >40 years or those with a history of low gonadotrophin response were given up to a maximum of eight ampules daily (administered in divided doses), with or without hMG (Repronex: Ferring, Tarrytown, NY, USA; Pergonal: Serono Laboratorones), either using an antagonist regimen or a microflare protocol. When at least two follicles reached a mean diameter of 16.5 mm and the estradiol17β level was >500 pg/ml, 10 000 IU hCG (Profasi; Serono) was administered intramuscularly followed 36 h later by transvaginal oocyte retrieval.

Luteal progesterone supplementation was initiated the day after oocyte retrieval and continued until 10 weeks in patients who became pregnant. Such luteal support was achieved by one of three regimens: (1) daily i.m. progesterone (50 mg), (2) daily vaginal gel [8% progesterone (Crinone; Wyeth-Ayerst, Madison, NJ, USA)]; or (3) three times daily vaginal progesterone suppositories (200 mg tid). Embryo transfer generally was performed with a Wallace catheter (Marlow/Cooper Surgical, Shelton, CT, USA). For difficult transfers, a Marrs No. 4 or Marrs No. 5 embryo transfer catheter (Cook Ob/Gyn, Spencer, IN, USA) was occasionally used.

Insemination and culture protocols

Oocytes were either inseminated 4–6 h after retrieval in groups [3–5 oocytes in 1.0 ml HF-10 (Sigma Aldrich, St. Louis, MO, USA) supplemented with 5% human serum albumin (InVitroCare Inc., Frederick, MD, USA)] overlaid with 1.0 ml oil in Falcon 3037 dishes (Becton Dickinson Labware, Franklin Lakes, NJ, USA) using 300 000 motile spermatozoa, or they underwent ICSI 3–5 h after retrieval. At the fertilization check, 16–18 h later, zygotes with 2 pronuclei (PN’s) were cultured individually in 25 µl microdrops of growth medium overlaid with 8 ml oil in Falcon 1007 culture dishes (Becton Dickinson Labware, Franklin Lakes, NJ, USA). From Days 1 to 3, one of three culture media was used [G1.3 (Scandinavian IVF Science/Vitrolife, Gothenburg, Sweden/Denver, CO; USA); Life-Global (IVF Online, Toronto, Canada); ECM (Sage/Cooper Surgical, Trumbull, CT, USA)]. All cultures were maintained at 37°C in a humidified atmosphere of 5% CO2 in air.

Embryo evaluations

All embryos were evaluated on Days 1, 2 and 3 with target evaluation times post-insemination or ICSI of 25 h for assessment of early cleavage in the afternoon of Day 1, 44 h for Day 2 evaluations, and 68 h for Day 3. Evaluation times were recorded for all embryos, with morphological characteristics scored and nominal data coded, as shown in Table I. On Day 1, embryos were graded for cell number, extent of fragmentation as assessed by the volume of the embryo occupied by fragments, and the extent to which the blastomeres were asymmetric (i.e. dissimilar in size and shape), and exhibiting nuclei. For the purposes of assessing the contribution of scoring the blastomere nuclei, the embryos were stratified into two groups: into those either having or not having a single nucleus in all blastomeres. On Day 3, embryos were similarly assessed for cell number, fragmentation and asymmetry, as previously described (Racowsky et al., 2003). Examples of various degrees of asymmetry and fragmentation are shown for 8-cell embryos in Fig. 1.

Table I List of embryonic features included in the study

<table>
<thead>
<tr>
<th>Day</th>
<th>Feature</th>
<th>Comments (including coding of nominal data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Time of evaluation</td>
<td>17.8–29 h after fertilization (target 25 h)</td>
</tr>
<tr>
<td>1</td>
<td>No. cells</td>
<td>Range: 1–6</td>
</tr>
<tr>
<td>1</td>
<td>Pronuclei</td>
<td>Code (day/ stage): 0 = 1 cell + 2PN; 1 = 1 cell + 0PN; 2 = 2 cells; 3 = &gt;2 cells</td>
</tr>
<tr>
<td>2</td>
<td>Time of evaluation</td>
<td>37.6–50.7 h (target = 44 h)</td>
</tr>
<tr>
<td>2</td>
<td>No. cells</td>
<td>Range: 1–11</td>
</tr>
<tr>
<td>2</td>
<td>Fragmentation*</td>
<td>Code: 0 = 0%; 1 = 1–9%; 2 = 10–25%; 3 = 26–50%; 4 = &gt;50%</td>
</tr>
<tr>
<td>2</td>
<td>Symmetry</td>
<td>Code: 1 = No asymmetry; 2 = Moderate asymmetry; 3 = Severe asymmetry</td>
</tr>
<tr>
<td>2</td>
<td>Distribution of nuclei</td>
<td>Code: 0 = one or more blastomeres without a single nucleus; 1 = all blastomeres with a single nucleus</td>
</tr>
<tr>
<td>3</td>
<td>Time of evaluation</td>
<td>60.2–77.7 h (target = 68 h)</td>
</tr>
<tr>
<td>3</td>
<td>No. cells</td>
<td>Range: 1–14</td>
</tr>
<tr>
<td>3</td>
<td>Fragmentation*</td>
<td>Code: 0 = 0%; 1 = 1–9%; 2 = 10–25%; 3 = 26–50%; 4 = &gt;50%</td>
</tr>
<tr>
<td>3</td>
<td>Symmetry</td>
<td>Code: 1 = No asymmetry; 2 = Some asymmetry; 3 = Severe asymmetry</td>
</tr>
</tbody>
</table>

*Due to small numbers in groups having scores of three and four, these groups were combined with those having a score of two.
Study dataset

Our database was screened to identify all cycles performed at Brigham & Women’s Hospital from 1 November 2003 through 31 March 2007 resulting in an embryo transfer on Day 3 without use of a gestational carrier, and in which all embryos were graded in the afternoon of the fertilization check (Day 1), and again on Day 2 and Day 3 (n = 1997). From this initial set of screened cycles, those having only one or two embryos transferred were further identified (n = 1257). This subset was then used to distinguish those cycles in which the developmental fate of each transferred embryo was definitively known (n = 972, with transfer of 1661 embryos). Developmental fate was classified as an embryo either failing to implant and give rise to a clinical sac, or implanting and developing into a viable fetus, detectable to at least 12 weeks of gestation. Accordingly, the single embryo transfers gave rise either to no detectable uterine pregnancy, or to a viable singleton, and the double embryo transfers resulted in either no detectable uterine pregnancy, or to a pregnancy comprised of viable dizygotic twins at 12 weeks.

Of the 1661 embryos identified with known developmental fate, 66 were subsequently excluded because of missing morphological data or because they either exhibited the rare condition of a single pronucleus (1PN) on the afternoon of Day 1 (despite having 2PNs at the fertilization check), or because they were transferred along with one of these unusually developing embryos. The final dataset comprised 1595 embryos, 564 of which resulted from ICSI and 1031 underwent assisted hatching with or without ICSI. In all, 269 embryos were transferred alone, and 1326 were transferred in double embryo transfers. A listing of the characteristics associated with the cycles from which the embryos were derived is given in Table II.

Biometrical considerations

Logistic regression models were fitted to the data since the response (presence or absence of a fetus) was a binary variable (Hosmer and Lemeshow, 2000). The following subsets were selected for initial model building, which we assumed are likely to have a major impact on the development of a fetus at 12 weeks. These were denoted:

D1: {eggDonor, cell counts, day1stage} on Day 1.
D2: {eggDonor, cell counts, fragmentation, symmetry scores and nuclear status} on Day 2.
D3: {eggDonor, cell counts, fragmentation and symmetry scores} on Day 3.

The day1stage incorporated cell number and the number of pronuclei on Day 1. The fragmentation and symmetry scores were classified as factor

Table II Characteristics of the cycles included in the study

<table>
<thead>
<tr>
<th>Feature</th>
<th>Comments (including coding of nominal data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal age</td>
<td>Range: 21.6–44.8 years</td>
</tr>
<tr>
<td>Source of embryos</td>
<td>Code: 0 = Patient; 1 = Donor</td>
</tr>
<tr>
<td>No. oocytes recovered</td>
<td>Range: 1–52</td>
</tr>
<tr>
<td>No. oocytes at metaphase II</td>
<td>Range: 1–44</td>
</tr>
<tr>
<td>Route of fertilization</td>
<td>Code: 0 = IVF; 1 = ICSI</td>
</tr>
<tr>
<td>Culture medium</td>
<td>Code: 0 = ECM; 1 = G1.3; 2 = Global</td>
</tr>
<tr>
<td>No. embryos that developed by Day3</td>
<td>Range: 1–35</td>
</tr>
<tr>
<td>Use of assisted hatching</td>
<td>Code: assisted hatching not used = 0; used = 1</td>
</tr>
<tr>
<td>No. embryos transferred</td>
<td>1 or 2</td>
</tr>
<tr>
<td>Outcome at 12 weeks</td>
<td>Code: 0 = no fetus developed; 1 = fetus(es) developed</td>
</tr>
</tbody>
</table>
variables, and therefore transformed to dummy variables using the lowest score as the reference score.

All models also included the age of the oocyte (i.e. the age of the patient in autologous cycles, and the age of the oocyte donor in cases of oocyte donation), and the source of the oocyte (autologous or donated), both of which we assumed to have a major impact on the outcome. In all, seven models were evaluated (Table III).

### Modeling techniques

Standard techniques for the analysis of clustered data are the generalized estimating equation (GEE) algorithm and the generalized linear-mixture algorithms. However, GEE models, as opposed to logistic regression, did not fit our data well (AUC < 0.6 in all considered models). Seven candidate models were assessed for discrimination and calibration using methodologies as recommended by Cook (2007). For discrimination, area under the receiving operator characteristic curve (AUC) was used as a measure of the ability to rank embryos according to probability of success, a high value indicating good ranking ability. For model calibration, the Hosmer–Lemeshow Goodness-of-Fit test was used which measures agreement between prediction and observed risks. When the P-value was > 0.05, the model was considered to be well calibrated.

The data were randomly split into a training set (80%) and a test set (20%). Each model was fit on the training set and then evaluated on a test set. To overcome sample-bias, the data were split and the models fitted 30 times. Acceptable models were defined as those (i) whose median AUCs were significantly higher than those of any other candidates, and (ii) whose median Hosmer–Lemeshow P-values for the training set stayed significantly above 0.05. The degree of significance was evaluated by a Wilcoxon test under two-sample (AUC) and one-sample (Hosmer–Lemeshow P-value) settings with P < 0.05 considered statistically significant. All possible pair-wise combinations of the AUCs for the seven models were tested using the Bonferroni adjustment (P-value cutoff is 0.05/7C2).

Since several of the models were not significantly different from one another, clinical considerations were also taken into account for the selection of the final model. Finally, the chosen model for embryo selection was constructed using the full dataset from 1595 embryos.

Fisher’s Exact test, and the Kruskal–Wallis Exact test were used where appropriate to assess differences between proportions, with P < 0.05 considered statistically significant.

### Results

#### Justification for pooling data from patients treated with IVF and ICSI

Distributions of the ages of patients treated by IVF and ICSI

The ages of the patients ranged from 21.6 to 44.8 years. The distributions of the ages of the patients whose ova were fertilized either by IVF or ICSI indicate that there was a tendency for patients treated with IVF to be slightly older (IVF: 35.22 ± 4.36 years, ICSI: 34.18 ± 4.37 years).

Allocation of the numbers of embryos transferred in patients treated with IVF and ICSI

There was a borderline significant difference between the proportion of IVF and ICSI cycles having one embryo transferred (26.7 versus 32.6%, respectively; P = 0.042). However, there was no significant difference between the percentages of fetuses produced by IVF and ICSI when embryos were either transferred singly (30.2 versus 24.5%, IVF versus ICSI; P = 0.334), or as a pair (22.4 versus 21.3%, IVF versus ICSI; P = 0.673). The data obtained using IVF and ICSI have therefore been pooled in subsequent analyses since no major differences were observed between the distributions of the maternal ages, the allocation of numbers of embryos transferred and the yield of 12-week-old fetuses.

#### Preliminary univariate analyses

Maternal age

The effect of maternal age was examined by fitting a linear logistic regression between the presence or absence of a fetus and the age of the woman. The estimated equation poorly fitted the data (Hosmer–Lemeshow test: P < 0.001). Adding the square of the age to the equation improved calibration (P = 0.037). The estimated equation is:

\[
\text{logit}P = -10.311 + 0.696 \times \text{age} - 0.012 \times \text{age}^2
\]

where P is the probability of a fetus being present at 12 weeks. A plot of this probability against maternal age is shown in Fig. 2. The probability of a pregnancy was highest when her age was about 28 years whether the embryos were transferred singly or in pairs (data not available for maternal age, therefore available only for patients whose ova were fertilized either by IVF or ICSI).
Autologous versus donated oocytes

The proportion of viable fetuses resulting from donated oocytes was significantly greater than that from use of autologous oocytes (44/126 [34.9%] versus 319/1469 [21.7%]; $P = 0.001$). Therefore, oocyte donation was included as a variable in model building.

Times when the embryos were observed

The mean times of observation on Day 1, Day 2 and Day 3 were 24.8 h (range 18.0–29.0 h), 44.2 h (range 37.6–50.7 h) and 68.6 h (range 60.2–77.7 h), respectively. The regressions of the logit of the proportion of fetuses on time for the three days were not significantly different from zero (Day 1: $b = -0.007$, $s_b = 0.063$; Day 2: $b = -0.00004$, $s_b = 0.053$; Day 3: $b = -0.039$, $s_b = 0.024$). Thus, the times when the embryos were scored had no significant effects on the outcomes and so were not included in the models.

Number of cells in embryos on Days 1, 2 and 3

On Day 1, those embryos at the 2-cell stage were most likely to form a fetus (30.7%), while those either still at the 1-cell stage (with or without their visible PN’s), or that had cleaved beyond the 2-cell stage showed lower viability (1-cell with 2pn: 12.1%, 1-cell without their visible PN’s), or that had cleaved beyond the 2-cell stage showed lower viability (1-cell with 2pn: 12.1%, 1-cell without their visible PN’s), or that had cleaved beyond the 2-cell stage showed lower viability (1-cell with 2pn: 12.1%, 1-cell without their visible PN’s), or that had cleaved beyond the 2-cell stage showed lower viability (1-cell with 2pn: 12.1%, 1-cell without their visible PN’s), or that had cleaved beyond the 2-cell stage showed lower viability (1-cell with 2pn: 12.1%, 1-cell without their visible PN’s), or that had cleaved beyond the 2-cell stage showed lower viability (1-cell with 2pn: 12.1%, 1-cell with >2-cell: 9.5%). On Day 2, maximum fetal development was associated with embryos having 4-cells (30.4%), while on Day 3, in those groups with robust numbers of embryos (i.e. >50), maximum fetal development was associated with 8-cell embryos (30.2%). Of note was development of fetuses from 12-cell embryos on Day 3, which was higher than that for 8-cell embryos (34.1 versus 30.2%). This observation may be due to sampling error of a small sample size.

To allow for the non-linear relationship between day of culture and the number of cells associated with fetal viability (Day 1: 2 cells, Day 2: 4 cells and Day 3: 8 cells), a derived variable was included in all models, defined by:

$$\text{Dev}(i) = \text{cell counts on Day}(i)$$

$$- \text{count with maximum yield of fetuses}(i), \text{where}\ i = 1, 2, 3.$$  

Fragmentation on Days 1, 2, 3

An increased level of fragmentation on either Day 2 or Day 3 was associated with a significant reduction in the number of fetuses that developed to 12 weeks [Day 2, Score 0 (28.0% fetuses) to Score 4 (0% fetuses); Day 3, Score 0 (29.4% fetuses) to Score 4 (0% fetuses)].

Asymmetry on Days 1, 2, 3

An increased level of asymmetry on either Day 2 or Day 3 was associated with a significant reduction in the number of fetuses that developed to 12 weeks [Day 2, Score 1 (28.1% fetuses) to Score 3 (9.1% fetuses); Day 3, Score 1 (31.1% fetuses) to Score 3 (5.0% fetuses)].

Multinucleate blastomeres on Day 2

The percentage of fetuses that developed from embryos was highest when all the blastomeres had a single nucleus (32.7%). The second largest percentage occurred when single nuclei could be seen in some, but not all, of the blastomeres (17.9%). Relatively smaller percentages of fetuses occurred when no nuclei were visible (9.0%), and when some blastomeres had more than one nucleus (9.9%). In the following multivariate linear regression analysis these data were reclassified in binary form where 0 was allotted to an embryo in which all blastomeres had a single nucleus and one was allotted to the rest.

Model selection

Figure 3 shows the AUC and Goodness of Fit $P$-values for both the training and test sets for each of the seven models investigated. There is close agreement between the results for the training (Fig. 3A, C) and test (Fig. 3B, D) sets. The median AUC of each of the seven models from the test sets was: D1 = 0.683; D2 = 0.729; D3 = 0.725; D1,2 = 0.725; D1,3 = 0.723; D2,3 = 0.739; and D1,2,3 = 0.737 (Fig. 3B). All seven candidate models showed adequate calibration, with the median Goodness-of-Fit $P$-value being significantly above the 0.05 threshold (Fig. 3D). The AUC for the Day 1 model was significantly lower than that of each of the other six models, indicating that the Day 1 model has the lowest discrimination ($P < 0.002$ for all 15 combinations). In contrast, the values of the AUCs for the remaining six models were not significantly different, indicating they had approximately equal discrimination ($P > 0.002$ for all 15 combinations).

From the six models with comparable discrimination, Day 2 and Day 3 models were selected for further study based on practical application in the clinic (see Discussion). Logistic regressions were then computed using the whole datasets for these two models (Table IV). A plot of observed versus expected outcomes grouped by fixed probability thresholds for the Day 3 model is shown in Fig. 4. The figure confirms that the calibration of the model is acceptable, although it slightly under estimates the
observed values. Similarly, calibration was acceptable for model Day 2 (data not shown). The AUCs for model Day 2 and Day 3 were 0.733 (95% CI, 0.701–0.765) and 0.732 (95% CI, 0.700–0.764), respectively.

**Discussion**

The purpose of an embryo scoring system in an IVF clinic is to aid in selection of the ‘best’ embryos for transfer. An acceptable system is one that orders the embryos reliably on a quantitative scale. We have derived such a system using logistic regression analysis on a robust dataset of embryos of known developmental fate. Resampling-based statistical modeling has guided the derivation of a model that shows that single day morphology scoring on Day 3, just before embryo transfer, provides similar predictive value to multi-day scoring. Furthermore, our results suggest that it is sufficient to score embryos for number of cells, fragmentation and symmetry on Day 3, and adjust for maternal age and oocyte donor status to select the best embryos for transfer. The variables and their regression coefficients used in the model are shown in Table IV (Day 3).

The analysis led to the derivation of another model that shows that embryos could be scored on Day 2 only. The variables and their coefficients for this model are also shown in Table IV (Day 2).

Calculation of AUC coupled with the logistic regression method to build prognostic models allows one to determine the discriminatory ability of each model. Using such an approach, several computational procedures could have been used for selection of variables, e.g. forward, backward and stepwise fitting. Since there are pitfalls associated with these procedures when they are used alone (Good and Hardin, 2003), it is generally recommended that relevant clinical knowledge be used to assist in the selection of the independent variables. We adopted this approach in the present analyses, in which all the models we examined were based mainly on *a priori* clinical knowledge. Similar model building using logistic regression has been successfully used to predict the probability of blastocyst formation from cultured cleavage stage pig embryos (Booth et al., 2007).

**Figure 3** The AUCs and the goodness of fit *P*-values fit for the seven putative models. (A) AUCs of the training set. (B) AUCs of the test set. (C) *P*-values of the training set. (D) *P*-values of the test set. A logistic regression model is considered to be acceptable if the *P*-value is greater than the *P* = 0.05 threshold (horizontal dashed lines in B and D).
Predictive value of the models

The probabilities that each embryo in our dataset will develop into a fetus at 12 weeks have been computed using the Day 3 model. The distributions of the individual probabilities associated with either the development of a fetus or failure to develop are shown in Fig. 5. Although the two distributions overlap, it is clear that higher probabilities are associated with the development of a fetus compared with those associated with the failure of a fetus to develop.

The predictive value of a model can be examined by the calculation of a classification table for a hypothetical situation determined by the selection of a cutoff probability. Such a table has been computed for the Day 3 model assuming a cutoff probability equal to 0.3 (Table V). Four predicted outcomes can be computed: correctly predicted occurrence of a fetus (true positive), correctly predicted absence of a fetus (true negative), false prediction of a fetus (false positive), and false prediction of a fetus failing to develop (false negative). The results, in this example, show that 68.6% of embryos are classified correctly.

The discrimination of a model is often assessed by the AUC. In the present context, the AUC is the probability that a randomly-chosen transfer yielding a fetus at 12 weeks is ranked higher than a randomly-chosen transfer that did not yield a fetus at 12 weeks (Pepe, 2003). The estimates of the AUC show that there is little to choose among all models examined except for Day 1. All values fall in the range 0.7 ≤ AUC < 0.8, which Hosmer and Lemeshow (2000) consider acceptable discrimination. The values for the Hosmer and Lemeshow statistics are also within acceptable limits for both training and test sets, indicating reasonable calibration. However, although the calibration of the models is numerically acceptable, we do not advocate using the probability estimate derived from the models as a proxy for the probability of the success of the transfer, as further work is needed to refine the assessment of calibration for these models.

### Table IV Regression coefficients for Day 2 and Day 3 models using the whole dataset

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Day 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-11.620</td>
<td>3.110</td>
<td>0.000</td>
</tr>
<tr>
<td>Age</td>
<td>0.668</td>
<td>0.197</td>
<td>0.001</td>
</tr>
<tr>
<td>Age squared</td>
<td>-0.012</td>
<td>0.003</td>
<td>0.000</td>
</tr>
<tr>
<td>Cells</td>
<td>0.391</td>
<td>0.153</td>
<td>0.011</td>
</tr>
<tr>
<td>Dev(cells)</td>
<td>-0.837</td>
<td>0.168</td>
<td>0.000</td>
</tr>
<tr>
<td>Fragmentation1 (1 versus 0)</td>
<td>0.031</td>
<td>0.182</td>
<td>0.865</td>
</tr>
<tr>
<td>Fragmentation2 (2 versus 0)</td>
<td>-0.303</td>
<td>0.212</td>
<td>0.153</td>
</tr>
<tr>
<td>isAllSingleDay2</td>
<td>0.439</td>
<td>0.149</td>
<td>0.003</td>
</tr>
<tr>
<td>isEggDonor</td>
<td>0.071</td>
<td>0.248</td>
<td>0.774</td>
</tr>
<tr>
<td>Symmetry2 (2 versus 1)</td>
<td>0.082</td>
<td>0.145</td>
<td>0.571</td>
</tr>
<tr>
<td>Symmetry3 (3 versus 1)</td>
<td>-0.301</td>
<td>0.296</td>
<td>0.309</td>
</tr>
<tr>
<td><strong>Day 3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-12.047</td>
<td>3.270</td>
<td>0.000</td>
</tr>
<tr>
<td>Age</td>
<td>0.706</td>
<td>0.198</td>
<td>0.000</td>
</tr>
<tr>
<td>Age squared</td>
<td>-0.012</td>
<td>0.003</td>
<td>0.000</td>
</tr>
<tr>
<td>Cells</td>
<td>0.206</td>
<td>0.055</td>
<td>0.000</td>
</tr>
<tr>
<td>Dev(cells)</td>
<td>-0.232</td>
<td>0.067</td>
<td>0.000</td>
</tr>
<tr>
<td>Fragmentation1 (1 versus 0)</td>
<td>0.138</td>
<td>0.194</td>
<td>0.477</td>
</tr>
<tr>
<td>Fragmentation2 (2 versus 0)</td>
<td>-0.622</td>
<td>0.221</td>
<td>0.005</td>
</tr>
<tr>
<td>isEggDonor</td>
<td>0.084</td>
<td>0.252</td>
<td>0.740</td>
</tr>
<tr>
<td>Symmetry2 (2 versus 1)</td>
<td>-0.264</td>
<td>0.136</td>
<td>0.052</td>
</tr>
<tr>
<td>Symmetry3 (3 versus 1)</td>
<td>-1.598</td>
<td>0.353</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Figure 4** Calibration plot of the observed proportion of embryos developing into fetuses at 12 weeks versus the prediction probability calculated using the whole dataset and the final Day 3 model. Prediction probabilities were grouped into six equal fixed sets.

**Figure 5** Distributions of probabilities associated with developing a fetus or not developing a fetus.
This means that the models presented here should be used to rank order the embryos for probability of success, but should not be used to provide patients with an estimated probability of success. In related work in other clinical domains, we show that discrimination of models is usually conserved across different settings, although calibration is not (Matheny et al., 2005; Ohno-Machado et al., 2006). This is particularly relevant in the assisted reproductive technology domain, given the baseline differences in success rates across IVF clinics and different characteristics of the population served by these programs. Therefore, it is necessary to assess the calibration of this model in other clinics. Until this is done, communicating the probability estimates from this model to patients or using its results as a basis for deciding how many embryos to transfer is not warranted.

The marginal increase in discrimination achieved using the model of all 3 days combined (Day 1, 2, 3: median AUC = 0.737) as compared with those using only Day 2 (median AUC = 0.733) or Day 3 (median AUC = 0.732) might suggest that multi-day scoring is desirable. However, any such slight gain to embryo selection must be weighed against possible detrimental effects of repetitive environmental perturbations to the embryos (light exposure, temperature and pH shifts etc.), caused by serial removal from the incubators. Moreover, the practicality of performing repeat observations within specific time-windows in a busy clinical IVF laboratory must also be taken into consideration. On balance, we propose that evaluations should be performed either on Day 2 or Day 3 alone, since such a paradigm essentially maximizes morphological selection of cleavage stage embryos, and is the simplest and potentially the safest approach for embryo health. Although available evidence suggests comparable pregnancy rates are achieved with either day of transfer in good prognosis patients, Day 2 transfer may improve the probability of pregnancy in poor responders (reviewed by Kyrou et al., 2008).

The considerable overlap in the distributions of probabilities associated with developing or not developing into a fetus (Fig. 5) gives rise to false positive and false negative predictions depending on the assumed cutoff probability (Table V). This overlap is likely due to a multitude of variables that may impact on whether an embryo will implant. Aside from the overall health of the gametes (as affected by parental age, genetic backgrounds and life-time environmental exposures), other factors including the ovarian stimulation, the culture conditions, the day of transfer, the degree of difficulty of the transfer procedure itself, and uterine receptivity issues all are likely to play a role.

Moreover, although there is a correlation between aberrant morphology and increased aneuploidy (Munne et al., 1995; Magli et al., 2007), normal cleavage stage morphology does not guarantee euploidy (Gianaroli et al., 1997; Munne, 2006).

### Evaluation of new methods

Considerable research is now being done on alternative non-invasive methods to morphological scores, such as amino acid analyses (Brison et al., 2007) and metabolomic (Seli et al., 2007) and proteomic (Katz-Jaffe et al., 2006) profiling that screen spent culture media. In the evaluations of these newly developing systems, it is important that the most accurate morphological systems are used. In view of the demonstrated efficacy of models based on morphology with an AUC > 0.7, the putative superiority of new methods should require that they result in an AUC > 0.7. Recently, Vergouw et al. (2008) have used classification tables to compare systems based on morphological scoring and metabolic profiling on Day 2 and Day 3. Their claim that metabolic profiling is more accurate than morphological scoring is not convincing since the accuracy of their morphological scoring systems (27.6 and 38.5%, Day 2 and Day 3, respectively), compared poorly to the accuracy of our Day 3 model (68.6%), and the accuracy achieved with use of their metabolomic profiling was not superior (69.0 and 53.6%, Day 2 and Day 3, respectively).

A widespread belief exists that all morphological scoring systems have low predictive value for embryo selection. Our work shows that this belief is not true. Since these methods are based on simple direct observations, it is important that they be used as standards for evaluating all new methods. When the assessment is in binary form (i.e. the presence or absence of a fetus or newborn), the control system should have an AUC > 0.7. Only by setting such objective benchmarks can new methods be compared.

### Conclusion

In conclusion, the findings reported in this paper indicate that single day morphological evaluation on either Day 2 or Day 3 provides similar predictive value to multi-day scoring. The regression coefficients derived from our statistical analyses (Table IV) can be applied prospectively to enable available embryos to be ranked for each patient prior to transfer. As such, we expect this algorithm to be applied in the clinic for the superior selection for transfer of the most developmentally competent embryo in each cohort, thereby resulting in an overall improvement in our viable pregnancy rates. We are currently planning a prospective evaluation of the efficacy of the Day 3 model in our clinic and eventually its validation in other clinics.

### Acknowledgements

We thank the embryology team at Brigham & Women’s Hospital for their expertise and commitment to excellence in the care of the gametes and embryos of our IVF patients. We extend particular appreciation to Gena Ratiu, M.D. for taking the embryo photographs.
Funding

This work was funded in part by grant RO1LM009520 from the National Library of Medicine, NIH (L.O.).

References


Fisch J, Sher G, Adamowicz M, Keskintepe L. The graduated embryo score predicts the outcome of assisted reproductive technologies better than a single day 3 evaluation and achieves results associated with blastocyst transfer from day 3 embryo transfer. Fertil Steril 2003;80:1352–1358.


Skiaidas CC, Jackson KV, Racowsky C. Early compaction on day 3 may be associated with increased implantation potential. Fertil Steril 2006;86:1386–1391.


Racowsky et al.


Submitted on November 25, 2008; resubmitted on April 30, 2009; accepted on May 6, 2009.