Comparison of pregnancy predictive models applied to women who received IVF/ICSI in Valencia (Spain), using ROC curves.

Ana Debón, Patricia Carracedo and Inmaculada Molina

Universitat Politècnica de València
Camino de Vera s/n, 46022 Valencia, Spain
{andeau@eio,
patcarga@posgrado.upv.es,
mamobo1@dca}upv.es
http://www.upv.es

Abstract. One of the main objectives of Assisted Reproductive Technology (ART) is the reduction of the number of multiple pregnancies to obtain the highest number of full term pregnancies. Therefore, ART has to select the best quality embryos for transfer. This paper proposes several generalized linear models for the prediction of the implantation potential of embryos taking into account the morphological variables of the embryos and clinical and cycle variables of the woman. Some morphological variables are considered factors for the evaluation of differences between the effect of each variable value on the embryo implantation potential. With this aim in mind, we considered only two embryo transfers for women under 35 who received IVF-ICSI treatment at the "La Fe" Hospital in Valencia from 2001 to 2006. Model validation was carried out using ROC curves. In short, we propose statistical tools which provide a clear framework for decision support for embryologists.

Keywords: Generalised linear models, ROC curve, Youden Index

1 Introduction

Delayed motherhood/fatherhood has contributed to the increasing number of couples having difficulty conceiving. For this reason, many couples have to go to Human Reproduction Units (HRU) in order to achieve a pregnancy using assisted reproduction techniques (ART) [13]. The increasing improvement in these techniques has resulted in good quality embryos able to develop and implant in the uterus, thus increasing pregnancy rates (PR). This use of ART has increased considerably in the past 20 years. However, it has also disturbingly increased the obstetric and perinatal risks involved in multiple pregnancies such as abortion, premature births, neonatal death and hypertension. This is why, nowadays one of the main objectives of ART is the reduction of the number of multiple pregnancies to obtain the highest number of full term pregnancies.

To increase pregnancy rates it is essential to identify which of a woman's embryos are more capable of correct implantation. There are several classifications such as those based on the morphological features of the embryo, the classification proposed by Association for the Study of Reproductive Biology (ASEBIR) based on several embryonic variables, and finally embryonic scores or statistical models which provide a score for each embryo representing its quality. Unlike previous classifications, statistical models are less subjective, decreasing the variability of embryo classification.

This paper proposes several generalized linear models (GLM) for predicting the implantation potential of embryos from the morphological variables of the embryos and clinical and cycle variables of the woman. Some morphological variables are considered factors for the evaluation of differences between the effect of each variable value on the embryo implantation potential as [3] but applied to an homogeneous population and with new variables. The validation of both models is carried out using Receiver Operating Characteristics (ROC) curves and the corresponding area under (AUC) ROC curves and the Youden index (J).

2 Materials and Methods

2.1 Data

The paper is a retrospective study of 5242 cycles of IVF-ICSI with transfers of one, two or three embryos on day 2 (second day after fertilization) in the Human Reproduction Unit at the University Hospital La Fe in Valencia from January 2003 to January 2006. This database takes into account a total of 26 variables y 5.803 registers after being cleaned to eliminate errors such as registers in which the number of embryos transferred does not corresponds to the number of cells or the grade.

In order to homogenize the studied population, we selected a sample on which to perform all subsequent analyzes. This sample consisted of women with transfers of two embryos who were less than 35 years old. The selection of this sample is due to the special features of women of an advanced age, particularly over 35 and transfers of one or three embryos. In addition only cases with proven two failed or correct implantations were selected.

2.2 Statistics

Statistical analysis was performed using the R environment for statistical computing [9] together with ROCR package [12]. ROCR is a package for evaluating and visualizing the performance of scoring classifiers using the statistical language R. This package makes it easy to use a Receiver Operating Characteristic (ROC) graphs as a way to evaluate concordance between models and real data.

In this study we used Generalised Linear Models to calculate the predictive value of the variables for the occurrence of ongoing pregnancy. GLM are an

extension of linear models that allow to use non-normal errors distributions (binomial, Poisson, gamma, etc ...) and non-constant variance. Unlike linear models, the GLM use a link function. That function linearize the relations between the response variable and the independent variables by transforming the response variable. These models allowed us to analyze binary data and logit models, with categorical predictors often called factors. The coefficients in logit models with categorical variables were used to study the differences in probabilities between different values in the independent variables. We have used the GLM to predict the embryo score obtained by means of fixed effects in order not to need any information about the correlation between embryos.

ROC curves provided an overall representation of accuracy, their implementation and GLM are well-described by [3]. We considered problems where the items could only belong to two classes and some classification models (or classifiers) that produce a continuous output (e.g., an estimate of failure probabilities) to which different thresholds may be applied to predict class. For each individual we have both the model prediction and the actual class. Given a classifier and a threshold, there are four possible outcomes, Table 1 shows the possibilities

Table 1. Results in table form

		True class	
		A	B
Predicted class	A	True positives	False positives
	B	False negatives	True negatives
		Total positives	Total negatives

The true positive rate (tp rate) of a classifier is estimated as

$$\text{tp rate} = \frac{\text{True positives}}{\text{Total positives}}.$$

The false positive (fp rate) rate of a classifier is estimated as

$$\text{fp rate} = \frac{\text{False positives}}{\text{Total negatives}}.$$

Additional terms associated with ROC curves are

$$\text{sensivity} = \text{tp rate} = \text{S}$$

and

$$\text{specificity} = 1 - \text{fp rate} = \frac{\text{True negatives}}{\text{Total negatives}} = \text{E}.$$

ROC graphs are two-dimensional graphs in which the tp rate is plotted on the Y axis and the fp rate is plotted on the X axis. If the test did not permit

discrimination between classes, the ROC curve was the diagonal joining the vertices from bottom left to top right. The accuracy of the test increased as the curve moved towards the upper left corner.

Our model was designed to show the impact of each value of categorical variables for the implantation potential of the individual embryo on day 2 of transfer. To evaluate the discriminative performance of the logistic model and to compare the classifiers, we wanted to reduce ROC performance to a single scalar value representing expected performance. Calculating the area under the ROC curve, the AUC was a portion of the area of the square unit, its value always being between 0 and 1, so random guessing procedures had an area of 0.5. Therefore, when the area under the ROC curve (AUC) increased, the classifier power also increased. Although AUC is the most commonly used global index of diagnostic accuracy the Youden Index [14] is also frequently used in practice [1, 11]. This index can be defined as $J = max_c sensivity(c) + specificity(c) - 1$, for any given threshold c, and ranges between 0 and 1. Complete separation of the distributions of the marker values for the diseased and healthy populations results in $J = 1$ whereas complete overlap gives $J = 0$. Obtaining this threshold c which corresponds to J is an attractive feature not present in the AUC. Once we have calculated the AUC for each model, we compare them using the pROC package [10]. The AUC are compared with statistical tests based on U-statistics theory or bootstrap [4, 2]. By default the delong method is used except for comparison of partial AUC, smoothed curves and curves with different direction in those cases bootstrap is used.

In addition, and first of all, we used Correspondence Analysis (CA), which is a multivariate statistical technique conceptually similar to Principal Component Analysis (PCA), but applies to categorical rather than continuous data. In a similar manner to PCA, it provides a means of displaying or summarising a set of data in two-dimensional graphical form. A detailed description of CA can be found in [6]. Recently, several related R packages have implemented this technique. For instance, the ca package by [5].

3 Results

With this selected sample, the first step is to carry out an exploratory study to find out which variables in the database are useful to predict pregnancy. With this aim in mind, we first analyzed the normality and homoscedasticity of the data using the Shapiro Wilk test and Levene test, respectively. The result of this analysis indicated that the data wre not normal and heteroscedastic, so that data were transformed in three ways: logarithmically and using square root and cube root, to avoid using non-parametric tests which have less power than parametric ones. Unfortunately, the transformed variables still fail both hypotheses so nonparametric methods had to be used. Finally from a total of 26 variables, 12 produced significant differences between women who became pregnant and those who did not. These variables were: number of blastomeres,

symmetry and fragmentation of the embryo (grade), woman diagnosis, estradiol[1], FSH3[2], FSH.total[3], embryo origin, oocytes, REM.capacit[4] and total number of fertilized embryos .

As we mentioned before, in this paper two GLM for predicting implantation potencial of embryos were compared:

- Model with embryo variables.- this model is based on the methodology proposed by [3].
- Model with embryo, maternal and clinical variables.- the extended model adding maternal and clinical variables to the embryo variables.

3.1 Model with embryo variables

In this model we used a GLM to calculate the predictive value of the categorical variables where we can use the Bernoulli distribution for binary response variable "correct" or "failed" *implantation* (i.e., 1 or 0). The basic aim of our analysis was to predict the way in which implantation potential varies by values of embryonic characteristics and therefore it was important to note that the predictors were independent discrete factors with values 2, 3, 4, 5 and 6 for number of blastomeras and 1, 2, 3 and 4 for grade. A p-value< 0.10 was considered statistically significant.

In order to obtain a parsimonious model, as a first step CA technique was applied to analyse if the combination of levels was possible. The CA was obtained using the ca package for R. The CA plot is shown in Figure 1, where the distances within grade and number of blastomere categories are quite large. Therefore, categories for the number of blastomeres and grade were not grouped.

Finally, 75% of the data were randomly selected to establish the models and the remaining 25% were saved exclusively for evaluation. There are many commonly used link functions for GLM, and their choice can be somewhat arbitrary. The Bernoulli and Binomial family links are logit, probit, cauchit, (corresponding to logistic, normal and Cauchy CDFs respectively) log and cloglog (complementary log-log), *Deviance* is a goodness of fit statistic for a model, it is calculated as 2 times the log-likelihood of the full model minus 2 times the log-likelihood of the model and $\nabla Deviance$ is the increment of *Deviance* of a model with respect to the previous one. To explore the changes in the number of terms in a model we consider the Mallows statistic C_p which penalizes the complexity of models as it increases with the number of parameters. Therefore, we had to choose the model with the lowest Deviance and C_p. From these results we could conclude that although there were no major differences in link, the best results were for the probit link fit. Therefore this link was used for the regression.

[1] female sex hormone responsible for normal sexual development of women and menstrual cycle regulation.
[2] Indicates the ovarian reserve that a woman has.
[3] the FSH hormone that is injected to women in treatment
[4] number of sperm obtained in the semen sample.

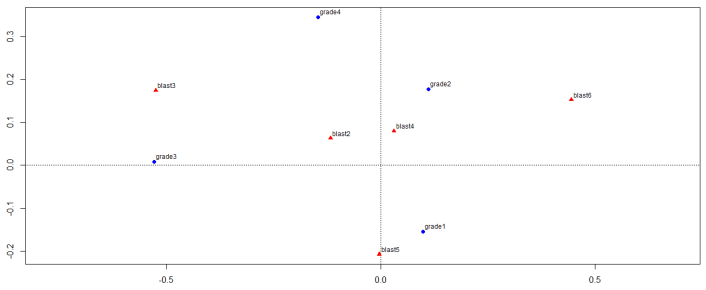

Fig. 1. Result of the Correspondence Analysis

After the fit of the probit GLM and from its results we were able to say that the "number of blastomeres" variable increase the probit of the implantation rate, specifically, the increase from 2 to 4 blastomeres increased the probit. The grade variable, however, significantly decreased this probit when moving from G1 to G3.

3.2 Model with embryo, maternal and clinical variables

The next step was to propose a new model, adding other variables in the database. The selection of those explanatory variables was performed by the "stepwise regression statistical". This technique was programmed manually in R environment [9]. From the results of the fit of this new model and as regards the values and sign of the factors: number of blastomeres and grade, we can observe that they are similar to the above model. In addition, these results are logical because with the increase of woman diagnosis IT decreases the value 2 being the worst. When ovarian reserve and the number of oocytes increases IT decreases , on the contrary when the total fertilized embryos decreases, IT increases.

- all the significant variables are coherent with non-parametic tests.
- the model is coherent with embryological practice.
- the women's age which is a very important variable is not in the model as it is correlated with FSH3.

3.3 Comparison of the models

Receiver Operating Characteristic (ROC) curves and the corresponding *AUC* were used to validate and to compare both models. Then, we considered the discriminating power of the score for each embryo whose implantation potential varied according to embryonic characteristics (number of blastomeres and grade) in the first model, and according with maternal and clinical variables in the

second model. The Figures 2 and 3 illustrate the ROC curves for the two models in both samples whose comparison allows us to assert that the second one assigns scores that discriminate better between women who are pregnant or not as the other model provides a curve closer to the diagonal and lower AUC.

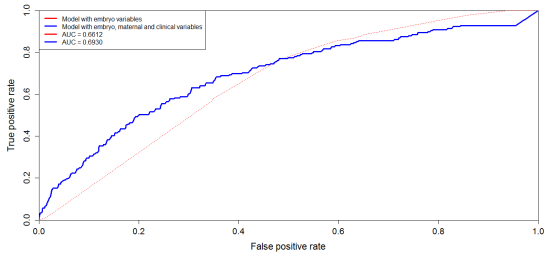

Fig. 2. Comparison of the ROC curves in modeling sample

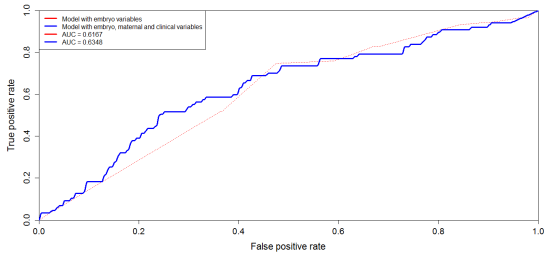

Fig. 3. Comparison of the ROC curves in validation sample

In addition, the AUC of the two models was compared using the pROC package [10] in both samples. The Table 2 shows the results, where it is observed that the p-value for both samples is highly significant, indicating that the models are valid. The AUC obtained for each model in the different samples, are within the average of the AUC. In both samples, we have used the bootstrap method because the curves have different direction.

After selecting the best model, we calculated the OOP scores in order to distinguish between pregnant and nonpregnant women. That threshold was calculated as the one whose vertical distance to the diagonal was greater, ie, one that maximizes the Youden index, as currently used in the medical field. Many authors advocate this approach [8]. To calculate this, we used OptimalCutpoints R-package [7].

Table 2. Table with pROC package results

Sample	**Method**	**D=$\frac{AUC1-AUC2}{\sigma}$**	**p-value**	**AUC roc model1**	**AUC roc model2**
Establish the models	Bootstrap	-4,56	5,02E-06	66,61	73,28
Validation	Bootstrap	-3,45	0,0005506	50,50	65,29

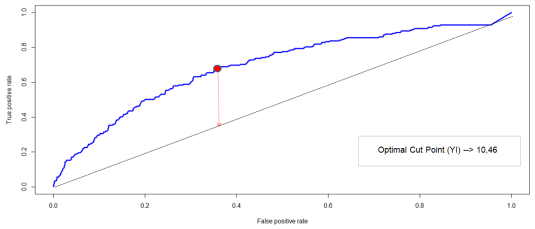

Fig. 4. OOP in the model with embryo, maternal and clinical variables in modeling sample

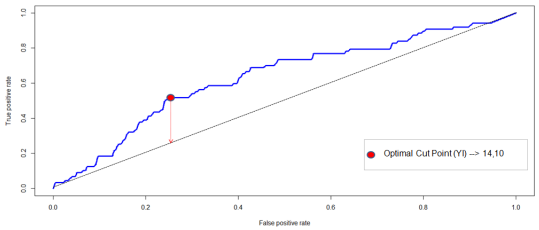

Fig. 5. OOP in the model with embryo, maternal and clinical variables in validation sample

4 Conclusions

In conclusion, our model provides a tool which facilitates decision making when choosing the best embryos for transfer. We have also proposed ROC curves as a graphical tool and the AUC and Youden Index as numerical values for validation and comparison of the different models.

This model can also be used in those databases in which, like ours, equality and symmetry variables are grouped into the grade variable.

References

1. Aoki, K., Misumi, J., Kimura, T., Zhao, W., Xie, T.: Evaluation of cutoff levels for screening of gastric cancer using serum pepsinogens and distributions of levels of

serum pepsinogen i, ii and of pg i/pg ii ratios in a gastric cancer case-control study. Journal of epidemiology/Japan Epidemiological Association 7(3), 143 (1997)
2. Carpenter, J., Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in medicine, 19(9), 1141-1164.
3. Debón, A., Molina, I., Cabrera, S., Pellicer, A.: Mathematical methodology to obtain and compare different embryo scores. Mathematical and Computer Modelling 57(5,6), 1380–1394 (2013)
4. DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837-845.
5. Greenacre, M., Nenadic, O., Nenadic, M.O.: Package ca (2012)
6. Greenacre, M.: La práctica del análisis de correspondencias. Fundación BBVA (2008), http://fbbva.es/TLFU/tlfu/esp/publicaciones/libros/fichalibro/index.jsp?codigo=300
7. Lopez-Raton, M., Rodriguez-Alvarez, M.X., Raton, M.M.L.: Package optimalcutpoints (2012)
8. Perkins, N.J., Schisterman, E.F.: The inconsistency of optimal cutpoints obtained using two criteria based on the receiver operating characteristic curve. American Journal of Epidemiology 163(7), 670–675 (2006)
9. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008), `http://www.R-project.org`, ISBN 3-900051-07-0
10. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., Mller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics, 12(1), 77
11. Schisterman, E. F., Perkins, N. J., Liu, A., Bondell, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology, 16(1), 73-81
12. Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T.: Rocr: visualizing classifier performance in r. Bioinformatics 21(20), 3940–3941 (2005)
13. Weinig, J.: Libro blanco sociosanitario: la infertilidad en España : situación actual y perspectivas. Imago Concept & Image Development (2011), http://books.google.es/books?id=oiMlMwEACAAJ
14. Youden, W.: Index for rating diagnostic tests. Cancer 3(1), 32–35 (1950)