Comparing Data Mining and Logistic Regression for Predicting IVF Outcome

59th Annual meeting of the American Society for
Reproductive Medicine (ASRM 2003)

J. R. Trimarchi, J. Goodside, L. Passmore, T. Silberstein, L. Hamel, L. Gonzalez

Objective

The most common question asked by IVF patients is "What are my chances of success?".
The answer to this difficult question typically considers patient age, FSH and infertility
diagnosis, however, many more parameters are known to impact the IVF cycle. We have
constructed mathematical models that provide a more accurate estimate of the outcome
of an IVF cycle.

Design

We utilized Quinlan's C5.0 decision tree data mining algorithm to retrospectively
investigate the predictive power of the 100 parameters that we track for each IVF cycle.
The parameters investigated include patient demographics, stimulation regime, response
properties, oocyte and embryo parameters and embryo transfer variables.   To validate
our findings from a statistical point of view we also constructed a statistical model based
on logistic regression.

Materials and Methods

The training data set consisted of 380 records of IVF data, each record described one IVF
cycle. We considered two experiments: a) We built a decision tree model with age as the
only predictor for pregnancy outcome. This gave us a model with roughly 56% accuracy
and is thought to closely resemble the kind of prediction currently in practice at IVF
clinics. b) We built a decision tree model, which considered about a dozen of the 100
available variables as true predictors. The model we obtained has an accuracy of about
75% in predicting pregnancy outcome.

We also computed a traditional statistic called concordance, based on multiple logistic
regression. In this approach we reduced the number of variables to about eight
statistically significant independent variables. Restricting further analyses to these
selected variables produced an association between predicted probabilities and observed
responses of about 74% concordant results, an increase in about 18% from a logistic fit
with just age as the predicting variable.

Results

We found a substantial increase in predictive power when considering more than just
age as a predictor of IVF outcome. The findings obtained from more traditional
statistical approaches seem to validate the results obtained by the data mining techniques
both in terms of accuracy and number of variables considered.

Conclusions

Consideration of many cycle specific parameters requires computer assistance and
decision tree models can predict IVF outcome with 75% accuracy on known outcomes of
cycles. More traditional statistical methods seem to confirm this. Mathematical models
should facilitate tailoring the IVF process to the individual patient and allow better
management of patient's expectations by predicting the probable outcome of the process.