Predicting IVF outcome
van Loendersloot, L.L.

Citation for published version (APA):
van Loendersloot, L. L. (2013). Predicting IVF outcome
Prediction models in in vitro fertilization; where are we?

L.L. van Loendersloot
S. Repping
P.M.M. Bossuyt
F. van der Veen
M. van Wely

Journal of Advanced Research 2013; in press: May 9
Chapter 5

ABSTRACT

Since the introduction of in vitro fertilization (IVF) in 1978, over 5 million babies have been born worldwide using IVF. Contrary to the perception of many, IVF does not guarantee success. Almost 50% of couples that start IVF will remain childless, even if they undergo multiple IVF cycles. The decision to start or pursue with IVF is challenging due to the high cost, the burden of the treatment, and the uncertain outcome. In optimal counselling on chances of a pregnancy with IVF, prediction models may play a role, since doctors are not able to correctly predict pregnancy chances. There are three phases of prediction model development: model derivation, model validation, and impact analysis. This review provides an overview on predictive factors in IVF, the available prediction models in IVF and provides key principles that can be used to critically appraise the literature on prediction models in IVF. We will address these points by the three phases of model development.
Prediction models in in vitro fertilization; where are we?

INTRODUCTION

Since the birth of Louise Brown in 1978, over 5 million babies have been born worldwide using in vitro fertilization (IVF) (1). The number of in vitro fertilization cycles has increased rapidly; in 2006, 458,759 cycles were reported in 32 European countries, 99,199 cycles in the USA and 50,275 cycles in Australia and New Zealand (2-4). The number of cycles is increasing each year even further.

The increase in IVF cycles is not caused by a sudden epidemic of infertility, but by increased access to IVF, and by an expansion of the indications for IVF. Initially, IVF was performed in couples with bilateral tubal occlusion (5). In 1992, intracytoplasmic sperm injection (ICSI) was first introduced and initiated in couples with severe male subfertility (6). Later on, IVF/ICSI was also applied in couples without an absolute indication for IVF, such as unexplained subfertility, cervical hostility, failed ovulation induction, endometriosis, or unilateral tubal pathology (7;8). The major difference between the original indication and the indications for which IVF is conducted nowadays is that the couples with bilateral tubal pathology or severe male subfertility have a zero chance of natural conception and completely depend on IVF/ICSI for a pregnancy, while couples with the newer indications are subfertile: they do have chances of natural conception, which may or may not be better than with IVF.

Despite the lack of evidence that IVF is effective in couples without an absolute IVF indication, IVF is often considered as a last resort for all subfertile couples regardless of the etiology of their subfertility (7-12). Contrary to the perception of many, IVF does not guarantee success; almost 38-49% of couples that start IVF will remain childless, even if they undergo six IVF cycles (13). Subfertile couples should therefore be well informed about the chances of success with IVF before starting their first or before continuing with a new IVF cycle. Based on a couple’s specific probability, one should decide whether the chances of success with IVF justify the burden, risks, and costs of the treatment. The threshold at which probability to start or to continue treatment may differ between different stakeholders, such as insurance companies, the tax payer, and the patients.

In optimal counselling on chances of a pregnancy after IVF, pregnancy prediction models may play a role, since doctors are not able to correctly predict pregnancy chances(14;15). Predictions made by clinicians on the basis of clinical experience or “gut-feeling” have only slight to fair reproducibility, indicating that these predictions are likely to be inaccurate (15).

The efforts to develop prediction models for IVF reflect the need for such models in clinical practice. This need can be explained by the inability of diagnostic tests to detect factors that indicate subfertility with near 100% certainty in patients. Accurate diagnostic tests would allow treatment to focus on specific factors (16). As IVF is currently used as an
empirical treatment and not as a causal intervention for a specific disorder, there is a strong need to distinguish between couples with a good and a poor prognosis (16). In the absence of randomized clinical trials, evaluating the effectiveness of IVF prediction models can be used to counsel couples.

The development of a prediction model can be divided into three phases: model derivation, model validation, and impact analysis (Fig. 1) (16;17). In the model derivation phase, predictors are identified, based on prior knowledge, and the weight of each predictor (regression coefficient) is calculated. In the model validation phase, the performance of the model, i.e. model’s ability to predict outcome is evaluated, and also the “generalizability” or “transportability” of the model is evaluated. The third and final phase consists of impact analysis. The impact analysis establishes whether the prediction model improves doctors’ decisions by evaluating the effect on patient outcome (16;17).

This review provides an overview on predictive factors in IVF, the available prediction models in IVF and provides key principles that can be used to critically appraise the literature on prediction models in IVF. We will address these points by the three phases of model development: model derivation, model validation, and impact analysis.

**Phase 1: model derivation**

*Identification of predictors*

Candidate predictors are variables that are chosen to be studied for their predictive performance. These can include subject demographics, clinical history, physical examination, disease characteristics, test results, and previous treatments (18). The identification of candidate predictors is preferably based on subject knowledge, on pathophysiological mechanisms, or the results of previous studies. Studied predictors should be clearly defined, standardized, and reproducible to enhance generalizability and application of study results to practice (18). Researchers frequently measure more predictors than can reasonably be analysed. When the number of predictors is much larger than the number of outcome events, there is a risk of overestimating the predictive performance of the model. To reduce the risk of false positive findings (predictors), at least 10 individuals having (developed) the event of interest are needed per candidate variable/predictor to allow for reliable prediction modelling (19).
A recent systematic review and meta-analysis on predictive factors in IVF evaluated nine predictive factors: female age, duration of subfertility, type of subfertility, indication for IVF, basal follicle stimulating hormone (bFSH), fertilization method, number of oocytes, number of embryos transferred, and embryo quality (20).

Female age is one of the most important prediction factors for success with IVF. Increasing female age was associated with lower pregnancy chances in IVF (OR 0.95, 95% CI: 0.94-0.96) (20). The decrease in fertility sets in after the age of 30 years, with a marked decline after 35 years for both spontaneous as IVF-induced pregnancies (20-23). The biological explanation for the declining chances to conceive with increasing female age most likely lies in the diminished ovarian reserve: the decrease in both quantity and quality of oocytes (24). Diminished ovarian reserve generally leads to a poor response to gonadotropin therapy and limits the possibility of a successful pregnancy (25).

Increasing duration of subfertility is known to be associated with a reduced possibility of natural conception (adjusted hazard rate 0.83; 95% CI 0.78-0.88) (7;26-30). In IVF, pregnancy rates were slightly lower in couples with a longer duration of subfertility (OR 0.99, 95% CI: 0.98-1.00), even after adjustment for age (20;23;31-33).

Although the meta-analysis did not find a significant association between type of subfertility (primary versus secondary subfertility) and pregnancy with IVF (unadjusted OR 1.04 95% CI: 0.65-1.43), two recent, large studies did find an association (20;31;33). A previous ongoing pregnancy or live birth, adjusted for factors such as age, substantially increases the likelihood of success with IVF (31;33).

Through the years, several studies have reported on the association between the indication for IVF and pregnancy with IVF without consistent results. These studies did not use the same reference categories making the interpretation of the data difficult. There is evidence for an association between tubal pathology and pregnancy with IVF. Women with tubal pathology alone had lower pregnancy chances compared to women with unexplained subfertility or other indications (23;31;34-36). On the other hand, another study suggested that women with tubal pathology had higher pregnancy chances after IVF compared with couples with unexplained subfertility, though not significantly (37). There is also evidence for an association between male subfertility and pregnancy with IVF. Although two studies (N=2,628 cycles) reported that couples with male subfertility have lower pregnancy chances than those with unexplained subfertility a very large cohort study (N=144,018 cycles) showed that couples with only male subfertility had increased pregnancy chances compared to couples with unexplained subfertility (31;35;36). Since these studies use different reference categories and different number of categories, it is not possible to compare these results optimally. For future studies and the development for prediction models, it would be useful to report every indication for IVF as a separate variable instead of combining all indications into one factor, to be able to compare all studies (20).
Basal FSH is an indirect estimate of ovarian reserve. A higher bFSH value was associated with lower pregnancy rates after IVF (OR 0.94; 95% CI: 0.88-1.00) (20).

Increasing number of oocytes was associated with higher pregnancy chances with IVF (OR 1.04, 95% CI: 1.02-1.07) (20). A recent large cohort study (N= 400,135) also showed a strong relationship between the number of oocytes and live birth rate with IVF. The association is not linear; the best chance of live birth is associated with approximately 15 oocytes (38).

Although the meta-analysis did not find a significant association between pregnancy chances with ICSI compared to IVF (OR 0.95, 95% CI: 0.79-1.14), a more recent large cohort study (N=144,018 cycles) reported higher chances with ICSI compared to IVF (OR 1.28, 95% CI: 1.25-1.31), even after adjusting for all relevant factors (OR 1.27, 95% CI: 1.23-1.31) (20;31).

The number of embryos transferred and embryo quality were associated with increased pregnancy chances (20).

Estimation of the regression coefficient
After identifying all potential predictors, a multivariable model can be constructed by regression analysis (logistic regression or proportional hazard analysis). To evaluate the quantitative effect of each predictor, the weight of each predictor is calculated by estimating the corresponding regression coefficient in a linear model.

Currently, over 21 papers have reported on the development and or validation of models for the prediction of pregnancy with IVF (Table I) (23;31-37;39-55).

Phase 2: model validation
The second phase in the development of a prediction model is the evaluation of the model performance, i.e. model validation. The performance of the model can be evaluated by calculating its discriminative capacity and the degree of calibration. Discrimination relates to how well a model can distinguish between patients with and without the outcome, i.e. discriminate between women who achieved pregnancy and those who did not. Discriminative capacity can be expressed by the area under the receiver operating characteristic curve (AUC), also known as the c-statistic. A model with a c-statistic of 0.5 has no discriminative power at all, while 1.0 would reflect perfect discrimination. Calibration relates to the agreement between observed outcomes and calculated probabilities, i.e. if we calculate a 30% probability of a pregnancy with IVF, the observed relative frequency of pregnancy should be approximately 30 out of 100 women. Calibration can be assessed by the Hosmer and Lemeshow goodness-of-fit test statistic. A Hosmer–Lemeshow statistics with a p-value above 0.05 implies that there is no significant miscalibration. In addition, calibration can also be assessed by comparing
the average calculated probabilities with the actual proportions in disjoint subgroups. The average calculated probabilities and actual proportions in each group can be plotted in a calibration plot. In case of perfect calibration, all points in a calibration plot are on the diagonal, the line of equality, and probabilities correspond perfectly to the actual proportions.

Table 1 | Characteristics of prediction models for pregnancy after IVF and IVF-eSET

<table>
<thead>
<tr>
<th>First author (year)</th>
<th>Inclusion of embryo characteristics</th>
<th>IVF-eSET</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Van Loendersloot et al. (2013)</td>
<td>Yes</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Nelson et al. (2011)</td>
<td>No</td>
<td>No</td>
<td>Live birth</td>
</tr>
<tr>
<td>van Weert et al. (2008)</td>
<td>No</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Lintsen et al. (2007)</td>
<td>No</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Verberg et al. (2008)</td>
<td>Yes</td>
<td>Yes</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Carrera-Rotlan et al. (2007)</td>
<td>No</td>
<td>No</td>
<td>Pregnancy</td>
</tr>
<tr>
<td>Ottoson et al. (2007)</td>
<td>Yes</td>
<td>Yes</td>
<td>Pregnancy</td>
</tr>
<tr>
<td>Fertlitsch et al. (2004)</td>
<td>No</td>
<td>No</td>
<td>Pregnancy</td>
</tr>
<tr>
<td>Hunault et al. 2002</td>
<td>Yes</td>
<td>Yes</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Stolwijk et al. (2000)</td>
<td>No</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Bancsi et al. (2000)</td>
<td>No</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Minaretzis et al. (1998)</td>
<td>Yes</td>
<td>No</td>
<td>Live birth</td>
</tr>
<tr>
<td>Commenges-Ducos et al. (1998)</td>
<td>Globel model: No</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Model for implantation: Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Templeton et al. (1996)</td>
<td>No</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Stolwijk et al. (1996)</td>
<td>Model A: No</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Model B: Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model C: Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bouckaert et al. (1994)</td>
<td>Yes</td>
<td>No</td>
<td>Pregnancy</td>
</tr>
<tr>
<td>Haan et al. (1991)</td>
<td>No</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Hughes et al. (1989)</td>
<td>No</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
<tr>
<td>Nayuda et al. (1989)</td>
<td>No</td>
<td>No</td>
<td>Ongoing pregnancy</td>
</tr>
</tbody>
</table>

The validation phase can be subdivided in internal validation (phase 2a) and external validation (phase 2b). With internal validation, the model's ability to predict the outcome in the group of patients in which it was developed is evaluated (reproducibility). Internal validation should be seen as validating the modeling process (56). Of the 21 papers reporting on IVF prediction model development, only 11 are also internally validated (23;31-35;37;40;45;49-51;53-55).
Before being able to use prediction models for clinical decision making it is not enough to demonstrate a reasonable or good performance after internal validation. Most models show too optimistic results, even after corrections from interval validation procedures. It is essential to confirm that any developed model also predicts well in a “similar but different” population outside the development set, i.e. external validation (generalizability). The more these populations differ from the development study, the stronger the test of generalizability of the model (57).

There are three different types of external validation, temporal validation, geographical validation, and domain validation. In temporal validation, the model is validated on new patients that are from the same center as the development set, but in a different time period (57;58). In geographical external validation, the model is validated on new patients from a different center as the development set (57;58). In domain validation, the model is validated on new patients that are very different from the patients from which the model was developed (57).

Of the 11 IVF models that went through internal validation, only four models have also been validated externally (32;33;37;45;49-51;53). One model was validated temporally, the model calibrated well both in the development set and in a separate validation set (33). Three models have been validated geographically, but only one model showed good calibration after validation (32;37;45;49-51;53). So at this moment, there is only one model that is generalizable to other clinics (37;45). All other models have to be geographically validated first before using the models in practice.

A prediction model often performs less well in a new group of patients than in the study group in which it was developed. This can be caused by differences in the case-mix between the development and validation population or by true differences between populations (58). Instead of simply rejecting the prediction model and develop or fit a new one, a better alternative is to update existing prediction models and adjust or recalibrate it to the local circumstances or setting of the validation set (57;58). As a result, the updated model is adjusted to the characteristics of new individuals. Several methods for updating prediction models are possible. Most often, differences are seen in the outcome frequency between the development and new validation set. This results in poor calibration of the model; predicted probabilities are systematically too high or too low. By adjusting the intercept (baseline risk) of the original model, calibration can be improved. Additional updating methods vary from adjustment of all predictor regression coefficients, adjustment of regression coefficients for particular predictor weight, to the addition of a completely new predictor or marker to the existing model (57;58).

As patient populations may shift during the years, the group of patients used for the development and validation of the prediction model may differ from the current patient population. Reproductive techniques may evolve during the years, new biomarkers with
Prediction models in in vitro fertilization; where are we?

predictive value may become available, and the prediction model should be regularly
updated and adapted to the new setting, so that predictions for future patients remain valid
and may even improve (58). IVF centers should therefore consider collecting their own
data in electronic databases, so that with accumulation of the number of IVF cycles over
time they can update the model with their own data.

Phase 3: impact analysis
The third and final phase in the evaluation of models is impact analysis; it establishes
whether the prediction model improves decisions, in terms of quality or cost-effectiveness
of patient care (17;57;58). This can be evaluated in one setting (phase 3a) or in varied
settings (phase 3b). Different study designs to evaluate the impact of a prediction model
are possible, such as comparing the outcomes between patients randomly assigned to
receive management guided by the prediction model and patients managed without the
prediction model (care-as-usual). A less valid alternative is to ask fertility specialists to
document therapeutic management decisions before and after being “exposed” to a model’s
predictions. None of the existing IVF prediction models has reached the impact analysis
phase yet.

DISCUSSION

As IVF can be stressful physically and emotionally and is not without health risks,
subfertile couples should thus be well informed about the chances for success with IVF
before each cycle. Unfortunately at this point, there are no randomized controlled clinical
trials comparing IVF with natural conception. Thus, the only way to counsel couples
properly is by model-based prognosis.

Over 21 papers have reported on the development and/or validation of prediction
models in IVF. Of these 21 papers, only two models had a good performance after
external validation. Impact analyses have not yet been performed for any of these
models. Future research should focus more on updating existing prediction models and
adjust or recalibrate them to the local circumstances or setting rather than developing
new prediction models. This way prediction models may strengthen evidence-based,
individualized decision-making and can contribute to a rational use of scarce resources.
REFERENCES


Prediction models in in vitro fertilization; where are we?


Prediction models in in vitro fertilization; where are we?


