Prospective validation of two models predicting pregnancy leading to live birth among untreated subfertile couples

Claudine C.Hunault¹,4, Joop S.E.Laven², Ilse A.J.van Rooij³, Marinus J.C.Eijkemans¹, Egbert R.te Velde³ and J.Dik F.Habbema¹

¹Department of Public Health and ²Department of Obstetrics and Gynecology, Division of Reproductive Medicine, Erasmus MC University Medical Center Rotterdam, PO Box 1738, 3000 DR Rotterdam and ³Department of Reproductive Medicine, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands

¹To whom correspondence should be addressed. E-mail: c.hunault@erasusmc.nl

BACKGROUND: Models predicting clinical outcome need external validation before they can be applied safely in daily practice. This study aimed to validate two models for the prediction of the chance of treatment-independent pregnancy leading to live birth among subfertile couples. METHODS: The first model uses the woman’s age, duration and type of subfertility, percentage of progressive sperm motility and referral status. The second model in addition uses the result of the post-coital test (PCT). For validation, these characteristics were collected prospectively in two University hospitals for 302 couples consulting for subfertility. The models’ ability to distinguish between women who became pregnant and women who did not (discrimination) and the agreement between predicted and observed probabilities of treatment-independent pregnancy (calibration) were assessed. RESULTS: The discrimination of both models was slightly lower in the validation sample than in the original sample which provided the model. Calibration was good: the observed and predicted probabilities of treatment-independent pregnancy leading to live birth did not differ for both models. CONCLUSIONS: The chance of pregnancy leading to live birth was reliably estimated in the validation sample by both models. The use of PCT improved the discrimination of the models. These models can be useful in counselling subfertile couples.

Key words: PCT/prediction model/subfertility/treatment-independent pregnancy/validation

Introduction

When counselling a subfertile couple, the decision to treat should be based on the pregnancy prospects without treatment of this specific couple and not on a uniform criterion, i.e. not having conceived within 12 months of unprotected intercourse. Treatments such as intrauterine insemination (IUI) or IVF should be proposed only to couples with a sufficiently low probability of treatment-independent pregnancy in order to avoid unnecessary medication and subsequent complications such as twin pregnancies, which is in itself associated with higher perinatal mortality rates and more long-term health and psychosocial sequelae (ESHRE Capri Workshop Group, 2000; Hansen et al., 2002; Stromberg et al., 2002; Jones, 2003; Moll et al., 2003). Treatment-independent pregnancy rates within 1 year after intake, i.e. after the diagnostic category has been definitively established, from 0 to 50% and more have been reported among couples with subfertility due to unexplained subfertility, mild male factor or cervical hostility (Eimers et al., 1994; Collins et al., 1995; Snick et al., 1997; Hunault et al., 2004). An accurate estimation of the chance of treatment-independent pregnancy for an individual couple is hence important, and may be provided by a prediction model.

We have previously developed two models to improve the prediction of treatment-independent pregnancy (Hunault et al., 2004). These models were based on three previous studies and are therefore called ‘synthesis models’. The population in which the two models were developed included couples consulting for various forms of subfertility (unexplained subfertility, subfertility due to cervical hostility or to a mild male factor), and referred by a general practitioner or by a gynaecologist. The first model includes the following predictors: the woman’s age, duration of subfertility, type of subfertility (primary or secondary), percentage of motile sperm cells and referral status of the couple. The second model includes the same predictors, plus the result of the best post-coital test (PCT). In clinical practice, such a model could be used to categorize a couple as having a poor, intermediate or good chance of conceiving without treatment. If the chance is poor, the couple should be advised to undergo treatment. If the chance is high, the couple should be encouraged to wait for treatment. If the chance is intermediate, the advice could be driven by the preferences of the couple concerning effectiveness, costs and risks of treatment.

The internal validity of the models has been found to be satisfactory, but an internally validated model can easily
produce poor predictions in future patients or in patients from other centres (Justice et al., 1999). The aim of the present study was to validate externally the two treatment-independent pregnancy prediction models, i.e. to assess whether these models predict well in a sample of subfertile patients different from the sample of patients used to develop the models.

Subjects and methods

Patients

This study was approved by the local institutional medical and ethical review boards, and written informed consent was obtained from all participants.

The standardized initial screening included the clinical examination of both partners, a (i.e. the first) semen sample analysed according to WHO criteria (World Health Organization, 1999), recording of a basal body temperature chart, a mid-luteal progesterone determination, a PCT, a transvaginal ultrasound and serum Chlamydia antibody testing. A hysterosalpingography or a laparoscopy with tubal patency testing was performed if Chlamydia antibody bodies were present or in the case of risk factors for tubal pathology (ectopic pregnancy or abdominal surgery history).

Three hundred and two couples from the Rotterdam and Utrecht University hospitals were enrolled prospectively in the study between January 1998 and August 2002. Inclusion criteria were: (i) woman’s age <40 years; (ii) duration of subfertility of ≥1 year; (iii) cycle duration >21 and <35 days; (iv) normal physical examination [no body shape and stature suggesting Turner’s syndrome, body mass index (BMI) <30 kg/m², normal secondary sexual characteristics, no abnormal findings on pelvic and gynaecological examination] and ultrasonography (no uterus abnormalities); (v) serum FSH concentrations within normal limits (1–10 IU/l); (vi) normal mid-luteal serum progesterone (≥28 nmol/l); and (vii) subfertility due to mild male, cervical or unexplained subfertility. Mild male factor was defined as a total motile count of at least 7 × 10⁶. Semen analysis was considered normal if sperm concentration was >14 × 10⁶/ml, if grade A progressive motility was >18% and if the percentage of normal morphology was >8% (strict Kruger criteria, Ombelet et al., 1997). The PCT was considered as positive if on average one progressively moving spermatozoa was found in at least six high power fields (World Health Organization, 1999). In case of a negative result, timing of the PCT was done using transvaginal ultrasound. Subfertility was attributed to cervical hostility if a correctly timed PCT revealed no progressive motile spermatoza in optimal cervical mucus in combination with normal semen samples, or if PCT was repeatedly negative regardless of the condition of the cervical mucus (World Health Organization, 1999). The diagnosis of unexplained subfertility was made when all investigations were normal. Couples with uni- and/or bilateral tubal disease, ovulatory disorder (abnormal serum progesterone in the mid-luteal phase) or endocrine disorders (abnormal prolactin or thyroid malfunction) or males with azoospermia were excluded. In summary, the inclusion and exclusion criteria of the population in which the models were validated were the same as those of the population in which the models were developed, except the semen criteria, which were stricter in the validation sample: in the development sample, only men with azoospermia were excluded, whereas men with severe male factor were also excluded in the validation sample.

All patient characteristics were collected prospectively: the woman’s age, duration of subfertility, type of subfertility (primary or secondary), percentage of motile sperm in the first semen analysis, result of the best PCT during the initial screening and referral status (whether the couple was referred by a general practitioner or by another gynaecologist). The following definitions were used. Duration of subfertility, the interval in years from discontinuation of contraceptive activities until registration at the fertility centre; primary subfertility, women who never conceived; secondary subfertility, subfertility after prior conception for the women; and live birth, living child at the time of hospital discharge after parturition. The number of observation months of couples was counted until either conception leading to live birth, or treatment was started, or because the study stopped before the end of their follow-up.

Analysis

Differences in couple characteristics between the validation sample and the original sample that provided the model were tested by Kruskal–Wallis test for continuous variables and χ²-test for categorical variables. The prognostic effects of the patient characteristics included in the model were studied in the validation sample and expressed as hazard ratios for live birth, using a multivariable model.

The synthesis models we aimed to validate are Cox models predicting the chance of treatment-independent pregnancy leading to live birth within 1 year after inclusion (Hunault et al., 2004). The model without PCT has been developed using data on 2459 couples obtained by pooling the data of three studies (Eimers et al., 1994; Collins et al., 1995; Snick et al., 1997). The model with PCT is based on the data of two studies (those of Eimers et al. and Snick et al.) since the PCT was not investigated in the third study (that of Collins et al.). The formulae of the models are given in the Appendix. The probability of live birth was calculated for each couple of the validation sample, according to both models.

The calibration and the discrimination of the models were assessed to test the validity of the model in the validation sample. Calibration refers to the agreement between predicted and observed probabilities of treatment-independent pregnancies, whereas discrimination is the model’s ability to distinguish between the women who became pregnant and those who did not.

Calibration was assessed graphically by plotting the observed 1 year live birth rate against the predicted one year live birth probability in a calibration plot (Miller et al., 1993). We statistically tested whether the mean predicted and observed probabilities of pregnancy leading to live birth were different. Furthermore, we tested whether the predictions were too extreme (too low estimates for low probabilities and too high estimates for high probabilities), and whether the observed and predicted ongoing pregnancy rates were systematically different (Harrell et al., 1996). The discriminative ability of the model was quantified by the c statistic, which is equivalent to an area under the receiver operating characteristic (ROC) curve. A c statistic ranges from 0.5 (no discriminative power) to 1 (perfect discrimination). The c statistic is the probability that from a random pair of women, the one with the highest predicted probability of treatment-independent pregnancy leading to live birth will be the first to succeed.

In order to assess and compare the clinical usefulness of the two models, the patients of the validation sample were grouped into three categories of predicted chances of treatment-independent pregnancy leading to live birth within 1 year, <20, 20–40 and ≥40%. Clinical usefulness of a model was expressed as the percentage of patients assigned by the model to the two extreme categories.

Calculations were performed using commercially available software packages (SPSS Inc., Chicago, IL, 1999 and S-plus 2000, MathSoft Inc., Seattle, WA, version 2000). A P-value <0.05 was considered to indicate statistical significance.
Results

Three hundred and two couples were included (213 patients from Utrecht and 89 couples from Rotterdam). The chance of pregnancy leading to live birth did not differ significantly between the Utrecht and Rotterdam clinics ($P = 0.15$). We pooled the two data sets into the ‘validation sample’ to assess the validity of the synthesis models. The couple characteristics of the development and validation samples are summarized in Table I. Women from the validation sample were older but their duration of subfertility was shorter compared with the women from the development sample. Secondary subfertility, normal PCT and referral by a general practitioner were more frequent in the validation sample. The time until treatment was much shorter in the validation sample, in which 71% of couples started a treatment within the first year after intake, compared with only 23% in the development sample.

The live birth rate estimate at 12 months did not differ significantly between the validation sample and the development sample. In the validation sample, couples with a normal PCT had a nearly four times higher chance of treatment-independent pregnancy leading to live birth than couples with an abnormal PCT after adjusting for the woman’s age, primary subfertility, duration of subfertility, motility and referral status [hazard ratio equal to 3.7, 95% confidence interval (CI): 1.09–12.7].

The $c$ statistic was 0.59 (95% CI: 0.46–0.73) and 0.63 (95% CI: 0.51–0.75) for the synthesis models without and with PCT, respectively. The predictions were not statistically too extreme (neither too low estimates for low probability patients, nor too high estimates for high probability patients), and no systematic difference was observed between observed and predicted pregnancy rates ($P = 0.13$ for the model without PCT and $P = 0.6$ for the model with PCT).

Table II shows that the model with PCT was clinically more useful than the model without PCT since the low and high prediction categories applied to 52% (18 and 34%) of the patients when using the model with PCT versus 36% (25 and 11%) when using the model without PCT. The two models tended to overestimate the probability of live birth in the category of predicted chances >40% because the estimate is only 36% (Table II). This is consistent with Figure 1.

Discussion

We assessed the validity of two models predicting the chance of pregnancy leading to live birth in untreated subfertile couples in a population different from the sample of patients used to develop the models. This study shows that the models were well calibrated, i.e. the predicted probabilities did not differ significantly from the observed probabilities. The model including the result of the PCT discriminated better between women who became pregnant and women who did not than the model without PCT ($c$ statistic equal to 0.63 and 0.59, respectively).

The discriminative ability was slightly lower in the validation sample than in the data of the three studies used to develop the models. In the latter, the $c$ statistic varied between 0.59 and 0.64 for the model without PCT and between 0.64 and 0.67 for the model with PCT after internal validation (Hunault et al., 2004). The lower $c$ statistics

---

Table I. Couple characteristics in the development sample ($n = 2459$ for the model without PCT and $n = 1398$ for the model with PCT) and in the validation sample ($n = 302$)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Median (upper and lower quartile)</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Development sample</td>
<td>Validation sample</td>
</tr>
<tr>
<td>Woman’s age (years)</td>
<td>29 (27–32)</td>
<td>32 (29–35)</td>
</tr>
<tr>
<td>Duration of subfertility (years)</td>
<td>2.5 (1.6–4)</td>
<td>2 (1.5–3)</td>
</tr>
<tr>
<td>Follow-up (months)</td>
<td>11 (5–27)</td>
<td>4 (2–7)</td>
</tr>
<tr>
<td>Total no. of sperm cells ($10^6$/ml)</td>
<td>42 (16–85)</td>
<td>55 (33–111)</td>
</tr>
<tr>
<td>Motile sperm cells (%)</td>
<td>47 (30–60)</td>
<td>54 (43–62)</td>
</tr>
<tr>
<td>Sperm cells with normal morphology (%)</td>
<td>55 (42–74)</td>
<td>12 (7–17)</td>
</tr>
<tr>
<td>% (no. of patients)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Secondary subfertility</td>
<td>22 (530)</td>
<td>37 (112)</td>
</tr>
<tr>
<td>Referral statusa</td>
<td>22 (530)</td>
<td>37 (112)</td>
</tr>
<tr>
<td>GP</td>
<td>30 (743)</td>
<td>61 (184)</td>
</tr>
<tr>
<td>Gynaecologist</td>
<td>68 (1675)</td>
<td>38 (114)</td>
</tr>
<tr>
<td>PCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>55 (769)</td>
<td>79 (238)</td>
</tr>
<tr>
<td>Abnormal</td>
<td>45 (624)</td>
<td>17 (52)</td>
</tr>
<tr>
<td>Missing</td>
<td>0 (5)</td>
<td>4 (12)</td>
</tr>
<tr>
<td>Treatment started &lt; 12 months</td>
<td>23 (574)</td>
<td>71 (215)</td>
</tr>
</tbody>
</table>

$a$Forty-one and four missing values for referral status, respectively, in the development and validation samples.
observed in the validation sample could be due to the fact that the validation sample is a more homogeneous group with patients having less extreme chances of pregnancy without treatment (predicted chance of treatment-independent pregnancy ranging between 5 and 68%, SD = 13 in the validation sample compared with predicted chance ranging between 1 and 75%, SD = 14 in the development sample).

PCT is an important predictor of treatment-independent pregnancy in this sample of patients. This result is interesting since the way in which the PCT is performed in one of the two study centres has changed in the last years. The effect of the result of the PCT in our model has been estimated using data from the study of Eimers et al. (1994) and that of Snick et al. (1997). In the study of Eimers et al., the PCT was performed in the fertility laboratory whereas it is currently performed by the clinicians (senior or junior residents). In the study of Snick et al., the PCT was performed by one of the four experienced gynaecologists of the peripheral hospital.

The prognostic power of the PCT has been established previously for couples with duration of subfertility < 3 years (Glazener et al., 2000), i.e. 80% of our validation sample. The repeated finding that the PCT is an important predictor suggests that the level of experience of the person performing the PCT does not have an effect.

Currently, various effective treatment modalities are available. In our validation sample, treatment was often started early, also for patients who still had a good chance of treatment-independent pregnancy, even in the centre with a long-standing history of use of clinical prediction models (the Utrecht clinic). Among the 27 patients with a predicted probability of ≥50% according to the model with PCT, 52% started a treatment within 6 months after intake (79% in the Utrecht clinic and 21% in the Rotterdam clinic). These 77 couples had a median duration of subfertility of 1.6 years, a median woman’s age of 29 years and a median sperm motility of 60%. Eighty-five percent of them were referred by a general practitioner and had a secondary subfertility. The PCT was normal in all cases. Because of the high percentage of treatment initiated within the first year, few treatment-independent live births were conceived in 1 year. The statistical power of Cox analysis is related to the number of events (45 treatment-independent pregnancies leading to live birth in this study) so the fact that no significant ‘lack of fit’ (calibration) of the model was detected does not mean that calibration was perfect. The calibration of the model should be confirmed in a study with a larger number of couples.

Could the use of the models improve the counselling of couples in comparison with the actual IUI and IVF guidelines of the Dutch Society of Obstetrics and Gynaecology (Dutch acronym: NVOG; www.nvog.nl)? According to these guidelines, IUI—and eventually IVF—treatments are offered to patients with unexplained subfertility according to the woman’s age and the duration of infertility. We categorized

---

**Table II.** Predicted probabilities and observed proportions of 1-year treatment-independent-pregnancy leading to live birth, using the synthesis models with and without PCT

<table>
<thead>
<tr>
<th>Predicted probability</th>
<th>n (%)</th>
<th>Observed 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model without PCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;20%</td>
<td>70 (25%)</td>
<td>7% 2–22%</td>
</tr>
<tr>
<td>20–40%</td>
<td>184 (65%)</td>
<td>33% 23–46%</td>
</tr>
<tr>
<td>&gt;40%</td>
<td>30 (11%)</td>
<td>36% 18–64%</td>
</tr>
<tr>
<td>Average prediction:</td>
<td>284 (100%)</td>
<td>32% 23–42%</td>
</tr>
<tr>
<td>Model with PCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;20%</td>
<td>52 (18%)</td>
<td>8% 3–22%</td>
</tr>
<tr>
<td>20–40%</td>
<td>135 (48%)</td>
<td>34% 20–53%</td>
</tr>
<tr>
<td>&gt;40%</td>
<td>97 (34%)</td>
<td>36% 24–52%</td>
</tr>
<tr>
<td>Average prediction:</td>
<td>284 (100%)</td>
<td>32% 23–42%</td>
</tr>
</tbody>
</table>

*Twelve patients with a missing value for PCT, four patients with a missing value for referral status and two patients with a missing value for outcome variable.

---

**Figure 1.** Calibration plots of the models: left panel without PCT and right panel with PCT. The squares correspond to the groups formed by pooling according to the predicted probabilities. The vertical lines are the 95% CIs of the observed 1 year live birth probabilities, estimated by Kaplan–Meier analyses.
the patients from the validation sample without missing values for the predictors of the models into two groups, patients who should be treated immediately and patients who should have an expectant management, according to the criteria of the Dutch IUI and IVF guidelines (see Table III). Within the group who should be treated immediately, 10% of the patients had a predicted probability of treatment-independent live birth >40% according to the model including PCT. In the group who should have expectant management, 11% of the patients had a predicted probability of treatment-independent live birth <20%. Moreover, about half of the patients fall into the intermediate class, in which patient preferences and counselling are particularly important. These findings suggest that use of the models may be valuable in clinical practice in addition to a guideline like the Dutch one. The patients with a predicted probability of <20% had a median duration of subfertility of 3 years, a median woman’s age of 33 years and a median sperm motility of 35%. Forty-eight percent of them were referred by a general practitioner and 19% had a secondary subfertility. The PCT was normal in 25% of the cases. The patients with a predicted probability of >40% had a median duration of subfertility of 1.7 years, a median woman’s age of 30 years and a median sperm motility of 54%. Eighty-one percent of them were referred by a general practitioner and 62% had a secondary subfertility. The PCT was normal in all cases.

Deciding whether or not a couple should be offered IUI or IVF treatment does not depend only on the probability of treatment-independent pregnancy. The probability of pregnancy with treatment is also important. If the latter is also low, starting treatment does not make sense.

If the models are used as a tool in counselling, the model with PCT is more useful than the model without PCT since the poor (<20%) and good (>40%) prognosis categories applied to more patients (52 versus 36%). The study has several implications for clinical patient practice. Only six readily available patients characteristics are necessary to use the model with PCT [woman’s age, duration of subfertility, type of subfertility (primary or secondary), referral status of the couple, progressive motility from the first semen analysis and result of the first correctly timed PCT]. The models apply to couples with subfertility due to unexplained reasons, cervical hostility and mild male factor. They have a broad basis of underlying patient populations and provide reliable predictions. Using these models would be useful for identifying those couples in which the treatment-independent chance of live birth is >40%. These couples should be strongly encouraged to refrain from any assisted reproduction treatment (ART) programme in the near future. These models might, furthermore, facilitate a more balanced choice of ART in those couples with lower chances of treatment-independent live birth.

Acknowledgements

We would like to thank Arie Verhoeff, Durk Berks and Lucienne Bax for their help in collecting the data in Rotterdam.

References


Submitted on November 25, 2004; resubmitted on January 27, 2005; accepted on January 27, 2005.
Appendix

The general formula of a Cox model is:
\[ S(t, x) = S_0(t) \exp^{(b_1 x_1 + \cdots + b_2 x_2)} = S_0(t)^{\exp(PI)} \]

The predicted probability \( (P) \) of treatment-independent pregnancy within 1 year after intake leading to live birth according to the synthesis model excluding the PCT result is:
\[ P = 100 \times (1 - 0.18^{\exp(PI)}) \]
where the prognostic index \( (PI) = -0.03 \times AGE1 - 0.08 \times AGE2 - 0.19 \times \text{duration of subfertility} - 0.58 \times \text{primary subfertility} + 0.008 \times \text{percentage of motile sperm} - 0.25 \times \text{tertiary–care couple}. \]

The formula of the synthesis model with PCT is:
\[ P = 100 \times (1 - 0.17^{\exp(PI)}) \]
and the prognostic index \( (PI) = -0.03 \times AGE1 - 0.06 \times AGE2 - 0.13 \times \text{duration of subfertility} - 0.44 \times \text{primary subfertility} + 0.008 \times \text{percentage of motile sperm} - 0.24 \times \text{tertiary–care couple} - 0.95 \times \text{abnormal PCT}. \]

AGE1 is the woman’s age if the age is \( \leq 31 \) years and \( 31 \) years if the age is \( > 31 \) years; AGE2 is the difference (woman’s age \( - 31 \) years) if the woman’s age is \( > 31 \) years and zero otherwise; a tertiary-care couple is a couple referred by a gynaecologist. Duration of subfertility is measured in years. For primary subfertility, tertiary couple and abnormal PCT, the value is 1 if true, 0 if not true.

The result of the PCT in the initial cycle was coded as abnormal when no forward-moving sperm cell was found in the whole mucus sample.