Human Reproduction vol.15 no.5 pp.1065–1068, 2000

External validation of the Templeton model for predicting success after IVF

J.M.J.Smeenk[1,3], A.M.Stolwijk[2], J.A.M.Kremer[1] and D.D.M.Braat[1]

[1]Department of Obstetrics and Gynaecology, University Hospital Nijmegen, PO Box 9101, NL-6500 HB Nijmegen, and [2]Department of Medical Affairs, University Hospital, Nijmegen, The Netherlands,

[3]To whom correspondence should be addressed

This study aimed to externally validate the prognostic model presented by Templeton in 1996 for live births resulting from IVF treatment. Data were used from the University Hospital, Nijmegen, The Netherlands, from March 1991 to January 1999. The predictive capacity of the model in our population discriminated between those women with a low probability of success and those with a relatively high probability. Despite these encouraging findings, our data show that implementation of the model in clinical decision-making remains difficult. The Templeton model is not applicable or usable in daily clinical practice, because the model did not give more information about the prognosis for the vast majority of the patients. Therefore, the search for better prognostic factors resulting in better predictive models should continue.

Key words: external validation/IVF/prediction/prognosis

Introduction

Modern medicine is more and more concerned with making choices from seemingly unlimited options. As a part of the decision making, for each individual, the physical, the psychological, as well as the financial costs should be weighed against the probability of success. Although it is practically impossible to predict the individual chance of a live birth in a couple accurately, prognostic models can help to encounter these matters in a rather objective way. They can also act as a convincing tool in individual counselling for both patients as well as physicians. It remains, however, to be seen how many patients refrain from treatment if their prognosticated chance is poor.

In the field of infertility, several authors have launched their models for the probability of pregnancy. Before any of these models can be implemented in clinical practice, good external validation is required (Stolwijk *et al*., 1998). The predictive accuracy of a prognostic model can be expressed by calibration and discrimination (Harrel *et al*., 1996). Calibration refers to the amount of bias in the predictions, while discrimination refers to the ability to separate patients with different outcomes. Unfortunately, prognostic models after IVF for the probability of pregnancy presented in the literature have often not been validated (Hughes *et al*., 1989; Haan *et al*., 1991; Templeton *et al*., 1996). One prognostic model was externally validated (Stolwijk *et al*., 1996); however, in another centre these tests proved that these models could not predict well. Templeton *et al*. (1996) developed a model to predict live birth after treatment with IVF using data from 26 389 women treated in all IVF centres in the UK. Although a model based on such a large population seems rather confirmative, it might not predict well in other populations. To examine the external validation, a model should be applied to other data than those upon which the model was based. To our knowledge, this external validation of the 'Templeton model' has not been done before. Therefore we started a retrospective study to validate the model and thereby estimate its clinical usability.

Materials and methods

Data were used from couples who underwent their first IVF treatment after March 1991 in the University Hospital Nijmegen, The Netherlands. All cycles, performed up to January 1999, were included. In concordance with the cycles included in the study of Templeton *et al*. (1996), cycles were excluded in which intracytoplasmic sperm injection (ICSI) was performed or in which donor gametes or frozen embryos were used. All cycles in which hormonal stimulation was initiated were included, regardless of whether follicle aspiration or embryo transfer was performed. Ovarian stimulation was most often performed by means of a long protocol of gonadotrophin-releasing hormone (GnRH) agonist, that was started on day 21 of the previous cycle, followed by human menopausal gonadotrophin (HMG). In general three embryos were transferred, but since January 1997 a maximum of two were transferred.

The predicted probability (P) of achieving a live birth after IVF was calculated using the Templeton the model:

$$P = \frac{100\times\exp(y)}{1 + \exp(y)}$$

Where y was defined as y = –2.028 + [0.00551×(age – 16)^2] – [0.00028×(age – 16)^3] + [i – (0.0690×no. of unsuccessful IVF attempts)] – (0.0711×tubal subfertility) + (0.7587×live birth after IVF) + (0.2986×previous pregnancy after IVF which did not result in a live birth) + (0.2277×live birth which was not a result of IVF) + (0.1117×previous pregnancy, not after IVF and which did not result in a live birth). Tubal subfertility and previous pregnancies were dichotomized in the model; 1 if applicable, 0 if not. The indicator 'i' was a value used to represent the infertility duration in years and was 0.2163 if the infertility duration was 1–3 years, –0.0839 if infertility duration was 4–6 years, –0.1036 if infertility duration was 7–12 years, and –0.4179 if infertility duration was ⩾13 years.

The Templeton model is based upon information from clinic forms which do not specify criteria for diagnosis (Craft and Forman, 1997). The variable 'diagnosis', as used in the model, is therefore the result of different work-ups and criteria. Furthermore, other variables were not specified at all. Because of these, we made a few assumptions to define the following variables in the model; (i) age: age of the woman at the specific IVF cycle; (ii) duration of infertility: duration of subfertility at the first IVF cycle; (iii) unsuccessful IVF attempts: the total number of previous IVF cycles in which no ongoing pregnancy was achieved (max. no. = 10); (iv) previous pregnancy not resulting in a live birth: spontaneous abortion or an ectopic pregnancy; (v) tubal pathology: tubal pathology exclusively; (vi) furthermore, because of limitations in the data available, we defined the predicted outcome: live birth. Because of incomplete follow-up we assumed for our calculations that all ongoing pregnancies, which are pregnancies that continued for at least 12 weeks after embryo transfer, resulted in live births.

We performed three external validations in which we intended to study the influence of different definitions by comparing the outcome of these three validations: (i) following the assumptions mentioned above; (ii) woman's age at the first IVF cycle (instead of at the specific IVF cycle); and (iii) tubal pathology exclusively or in combination with one or more other subfertility diagnoses (male factor, endometriosis, or cervical factor) (instead of tubal pathology exclusively).

We evaluated the predictive performance of the model by means of, firstly, the c index, which indicates the overall discriminative performance (Harrell *et al.*, 1982, 1996), and secondly, compared observed and predicted proportions of success for groups with a low probability (<5%, <10%) and a high probability (⩾20%). We presented predicted proportions with mid-*P* exact 95% confidence intervals (CI) (Vollset, 1993). The c index (number of concordant pairs + 0.5×the number of tied pairs/total number of pairs) can be interpreted as the probability of a correct prediction for a random pair that comprises a woman with an ongoing pregnancy and a woman without an ongoing pregnancy. A c index of 0.5 indicates that the predictions made for the whole population are bad; such a prediction is comparable to a flip of a coin. A c index of 1 indicates the ability to make perfect predictions.

Results

In total the data of 1292 couples who started a first IVF treatment since March 1991 in the University Hospital Nijmegen, The Netherlands, were used. Up to January 1999 they underwent 2756 IVF cycles. Of 35 couples who underwent 75 cycles, the duration of infertility was unknown; of two of them the subfertility diagnosis was also unknown. Of another three couples, who underwent five cycles, it was unknown whether tubal pathology was present. Of one couple who underwent two cycles it was unknown whether any non-IVF live births were present. After excluding data of these couples, we could use the data of 1253 couples who underwent 2674 IVF cycles for external validation of the Templeton model.

The mean age of the women at the beginning of treatment was 32.8 years (SD = 4.0; range 22–44; median 33 years) and the mean duration of infertility was 3.7 years (SD = 2.5; range 1–21; median 3 years). The mean number of previous unsuccessful IVF attempts was 0.8 (SD = 1.0; range 0–6; median 0 unsuccessful attempts). In the validation, 7% of the cycles were preceded by at least one previous live birth after IVF, 6% by at least one previous IVF pregnancy not resulting in a live birth, 13% by at least one live birth (excluding IVF births) and 22% by at least one pregnancy not resulting in a live birth (excluding IVF pregnancies). From all cycles that were used in the validation, 47% were first cycles, 29% were second cycles, 16% were third cycles, 5% were fourth cycles and 2% were of a higher rank (range 5–8). The distribution of indications for treatment, also important to test Templeton's model, is shown in Table I.

Table I. Indication characteristics of the couples at the start of the first IVF cycle (n = 1253)

Indication for treatment	Frequency (%)
Tubal pathology exclusively	295 (23.5)
Male factor exclusively	300 (23.9)
Tubal pathology and male factor	60 (4.8)
Other reasons[a]	338 (27.0)
Idiopathic	260 (20.8)

[a]Other reasons include hormonal, endometriosis, cervical factor, or a combination with tubal pathology or male factor

In the first validation (A) in which we used the assumptions mentioned above, we found a c index of 0.629. In the second validation (B), where another way of defining the woman's age was investigated, the c index was 0.632. In the third validation (C), where we looked upon the effect of another way to define the diagnosis, we found a c index of 0.628. Using the assumptions of validation A, we calculated the predicted proportions. In Table II, we present for each group of patients within a specific range of predictions (e.g. 0 to <5%) the observed proportion of ongoing pregnancies with the 95% confidence interval. The observed proportions increase from 0.0% in the group with a predicted probability of 0–<5% to 37.0% in the group with a predicted probability of ⩾30%.

In our population, the women with a low predicted chance (<10%) are relatively old (34–45 years) and never had a live birth after IVF treatment. Of these, the younger ones (34–36 years) all had a history of infertility of ⩾4 years. The women in our population with a fairly high predicted chance (⩾20%) generally were younger (66% were younger than 34 years) and most of them (86%) had a history of infertility of 1–3 years. The group with a high predicted chance (⩾30%) was characterized by women who all had a previous live birth after IVF.

Discussion

The comparison of the predicted and observed chances of success (Table II) shows that the model seems to be able to identify the women with a low chance and the women who have a high chance of achieving a live birth. Our c indices of ~0.6, however, suggest a poor predictive performance of the Templeton model. In general, ~13.9% of the IVF cycles will be successful, according to Templeton. Without any information about a patient, this will be the prior probability of success. A prognostic model is useful if it changes this prior probability in an accurate way. In the population, 76%

Table II. Predicted and observed probability of an ongoing pregnancy and percentage of ongoing pregnancies observed during an IVF treatment cycle in the first validation.

	Predicted probability (%)						
	0 to <5	5 to <10	10 to <15	15 to <20	20 to <30	⩾30	Total
No. of cycles	50	370	1116	924	187	27	2674
No. of ongoing pregnancies	0	34	179	180	67	10	470
Percentage of ongoing pregnancies (95% CI)	0.0 (0.0–5.8)	9.1 (6.5–12.5)	16.0 (14.0–18.3)	19.5 (17.0–22.1)	35.8 (29.2–42.9)	37.0 (20.6–56.2)	17.6 (16.2–19.1)

Table III. Women's ages and the resulting regression coefficients (Rc)

Age	Rc	Age	Rc	Age	Rc
20	0.07024	28	0.30960	36	–0.03600
21	0.10275	29	0.31603	37	–0.16317
22	0.13788	30	0.31164	38	–0.31460
23	0.17395	31	0.29475	39	–0.49197
24	0.20928	32	0.26368	40	–0.69696
25	0.24219	33	0.21675	41	–0.93125
26	0.27100	34	0.15228	42	–1.19652
27	0.29403	35	0.06859	43	–1.49445

had a (posterior) predicted probability of 10 to <20%. Such a prediction does hardly change their prior probability. Thus for the main proportion of patients the model of Templeton showed no use.

In the model the relative importance of the presented factors can be deducted from the parameter estimates resulting from the multiple logistic regression model. The 'duration of infertility' as well as 'previous pregnancies' play an important role (in the latter all applicable variables are multiplied by the regression coefficient), whereas the influence of the woman's age is not so obvious. Therefore we made a calculation of the relative effect of the woman's age. For this purpose we used the formula presented in the model:

$$0.00551\times(\text{age} - 16)^2 - 0.00028\times(\text{age} - 16)^3$$

From Table III it becomes clear that 29 years is the most favourable age to achieve a live birth after IVF, with the likelihood rapidly decreasing as the patient becomes older and that the relative positive influence of low age decreases in younger women. Although the parameters used by Templeton *et al*. all contribute to the predictive capacity of the model, age still is a very important predictor. We could not find remarkable differences between the results of our original validation (A) and our second validation (B), suggesting that there is no significant influence of the definition of the woman's age. This was expressed by the virtually unchanged c index (from 0.629 to 0.632).

In our third validation (C) we compared the predictive value of tubal pathology in combination with other diagnoses as subfertility diagnosis with 'tubal pathology exclusively'. The c index hardly changed (from 0.629 to 0.628). Therefore we concluded that the exact definition of 'tubal pathology' as subfertility diagnosis plays a minor role. This can be explained by the low regression coefficient for 'tubal reasons for infertility' in the 'Templeton model' (–0.0711).

In our assumptions we chose to use ongoing pregnancy as our endpoint instead of live birth, because the follow-up of pregnancies was not accurate enough. Data from our own clinic show that from July 1991 to December 1997, 506 ongoing pregnancies resulted in 482 live births (95%). The predicted probabilities for an ongoing pregnancy will therefore overestimate the expected probabilities of a live birth. We observed for the entire population that in 17.6% of the started cycles an ongoing pregnancy was achieved. The model by Templeton *et al*. predicted that in 14.4% (95% CI = 13.1–15.7%) of the started cycles a live birth would be achieved. Craft and Forman (1997) pointed out that Templeton reported an unexplained infertility incidence of >30%, which they felt was very high, considering that patients were treated in tertiary fertility referral centres. Our data show a considerably lower percentage (20.8%) of unexplained infertility cases. Last but not least, the original study revealed big differences between the contributing centres, which could attribute to the poor reproducibility of the model.

The question arises whether the development of a better model is possible or not. Other promising predictive factors may increase the predictive value of a model, as pointed out by Craft and Forman (1997). For instance, the basal FSH (Sharif *et al*. 1998), or day 3 oestradiol (Smotrich *et al*., 1995) values showed better predictive value than age alone. Inhibin is regarded to be another promising predictor of the outcome of IVF (Seifer *et al*., 1997; Lindheim *et al*., 1998). Moreover, since new techniques and medication influence the results of assisted reproductive technologies, a prognostic model has a limited lifetime and needs constant adaption.

In conclusion, the model presented by Templeton *et al*. based upon an unrivalled large data set, seems to be able to discriminate between a group of women with a very low probability of achieving success after IVF and those with a very high probability. However, for the majority of women, the application of the Templeton model did not give any more certainty, because their prior and posterior probabilities hardly differed.

References

Craft, I. and Forman, R. (1997) Analysis of IVF data. [Letter.] *Lancet*, **349**, 284.

Haan, G., Bernardus, R.E., Hollanders, J.M.G., *et al.* (1991) Results of IVF from a prospective multicentre study. *Hum. Reprod.*, **6**, 805–810.

Harrell, F.E., Califf, R.M., Pryor, D.B., *et al.* (1982) Evaluating the yield of medical tests. *J. Am. Med. Assoc.*, **247**, 2543–2646.

Harrell, F.E., Lee, K.L. and Mark, D.B. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.*, **15**, 361–387.

Hughes, E.G., King, C. and Wood, E.C. (1989) A prospective study of prognostic factors in *in vitro* fertilization and embryo transfer. *Fertil. Steril.*, **51**, 838–844.

Lindheim, S.R., Chang, P.L., Vidali, A. *et al.* (1998) The utility of serum progesterone and inhibin A for monitoring natural-cycle IVF-ET. *J. Assist. Reprod. Genet.*, **15**, 538–541.

Seifer, D.B., Lambert-Messerlian, G., Hogan, J.W. *et al.* (1997) Day 3 serum inhibin-B is predictive of assisted reproductive technologies outcome. *Fertil. Steril.*, **67**, 110–114.

Sharif, K., Elgendy, M., Lashen, H. and Afnan, M. (1998) Age and basal follicle stimulating hormone as predictors of *in vitro* fertilisation outcome. *Br. J. Obstet. Gynaecol.*, **105**, 107–112.

Smotrich, D.B., Widra, E.A., Gindoff, P.R. *et al.* (1995) Prognostic value of day 3 estradiol on *in vitro* fertilization outcome. *Fertil. Steril.*, **64**, 1136–1140.

Stolwijk, A.M., Zielhuis, G.A., Hamilton, C.J.C.M. *et al.* (1996) Prognostic models for the probability of achieving an ongoing pregnancy after *in vitro* fertilization and the importance of testing their predictive value. *Hum. Reprod.*, **11**, 2298–2303.

Stolwijk, A.M., Straatman, H., Zielhuis, G.A. *et al.* (1998) The search for externally prognostic models for ongoing pregnancy after *in vitro* fertilization. *Hum. Reprod.*, **13**, 3542–3549.

Templeton, A., Morris, J.K. and Parslow, W. (1996) Factors that affect outcome of in-vitro fertilization treatment. *Lancet*, **348**, 1402–1406.

Vollset, S.E. (1993) Confidence intervals for a binomial proportion. *Stat. Med.*, **12**, 809–824.

Received on June 21, 1999; accepted on January 11, 2000