Comparison of two models predicting IVF success; the effect of time trends on model performance


1Department of Public Health, Erasmus MC University Medical Centre, Rotterdam, The Netherlands 2Department of Obstetrics and Gynaecology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands 3Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, The Netherlands

*Correspondence address. E-mail: e.r.tevelde@ziggo.nl

Submitted on April 18, 2013; resubmitted on August 19, 2013; accepted on September 16, 2013

STUDY QUESTION: How well does the recently developed UK model predicting the success rate of IVF treatment (the 2011 Nelson model) perform in comparison with a UK model developed in the early 1990s (the Templeton model)?

SUMMARY ANSWER: Both models showed similar performance, after correction for the increasing success rate over time of IVF.

WHAT IS KNOWN ALREADY: For counselling couples undergoing IVF treatment it is of paramount importance to be able to predict success. Several prediction models for the chance of success after IVF treatment have been developed. So far, the Templeton model has been recommended as the best approach after having been validated in several independent patient data sets. The Nelson model, developed in 2011 and characterized by the largest development sample containing the most recently treated couples, may well perform better.

STUDY DESIGN, SIZE, DURATION: We tested both models in couples that were included in a national cohort study carried out in the Netherlands between the beginning of January 2002 and the end of December 2004.

PARTICIPANTS/MATERIALS, SETTING, METHODS: We analysed the IVF cycles of Dutch couples with primary infertility (n = 5176). The chance of success was calculated using the two UK models that had been developed using the information collected in the Human Fertilisation and Embryology Authority database. Women were treated in 1991–1994 (Templeton) or 2003–2007 (Nelson). The outcome of success for both UK models is the occurrence of a live birth after IVF but the outcome in the Dutch data is an ongoing pregnancy. In order to make the outcomes compatible, we used a factor to convert the chance of live birth to ongoing pregnancy and use the overall terms ‘success or no success after IVF’. The discriminative ability and the calibration of both models were assessed, the latter before and after adjustment for time trends in IVF success rates.

MAIN RESULTS AND THE ROLE OF CHANCE: The two models showed a similarly limited degree of discriminative ability on the tested data (area under the receiver operating characteristic curve 0.597 for the Templeton model and 0.590 for the Nelson model). The Templeton model underestimated the success rate (observed 21% versus predicted 14%); the Nelson model overestimated the success rate (observed 21% versus predicted 29%). When the models were adjusted for the changing success rates over time, the calibration of both models considerably improved (Templeton observed 21% versus predicted 20%; Nelson observed 21% versus predicted 24%).

LIMITATIONS, REASONS FOR CAUTION: We could only test the models in couples with primary infertility because detailed information on secondary infertile couples was lacking in the Dutch data. This shortcoming may have negatively influenced the performance of the Nelson model.

WIDER IMPLICATIONS OF THE FINDINGS: The changes in success rates over time should be taken into account when assessing prediction models for estimating the success rate of IVF treatment. In patients with primary infertility, the choice to use the Templeton or Nelson model is arbitrary.

Key words: IVF / prediction models / comparative performance / external test data / external validation
Introduction

Prior to the late 1970s almost no effective treatment for infertility was available. This situation changed after the birth of Louise Brown in 1978, the first child conceived by IVF. Due to improvements in stimulation regimes and innovations of the technique, results improved and IVF became a successful technique for the treatment of infertile couples. The greatest break-through came with the introduction of ICSI—a variant of IVF—in the early 1990s (Palermo et al., 1992). The ICSI technique made it possible to treat couples who were infertile because of serious sperm impairment. IVF success rates per treatment cycle increased in the UK from 14.0% in 1991 to 24.1% in 2009 (HFEA register 2011 (http://www.hfea.gov.uk/docs/2011-11-16), Fig. 1).

Nowadays IVF is applied not only in couples with sterility as a result of bilateral tubal occlusion as in the early days but also in couples with reduced fertility such as mild to moderate male infertility, cervical hostility or unexplained infertility. The probability of achieving a spontaneous pregnancy under these conditions may vary from zero to almost normal range (Hunault et al., 2004; Habbema et al., 2009; Brandes et al., 2010). Being able to compare the probability of natural conception to that of conception after IVF is advantageous for such couples because unnecessary and burdensome treatment may be avoided when the probability of natural conception appears to be still reasonable. By contrast, when the probability of natural conception is low compared with conception after IVF, the couple can be advised to proceed to IVF without any delay.

Several prediction models have been developed to estimate the probability of a spontaneously occurring pregnancy (Eimers et al., 1994; Collins et al., 1995; Snick et al., 1997; Hunault et al., 2004; Van der Steeg et al., 2007) and to estimate the probability of a pregnancy after IVF (Stolwijk et al., 1996, 1998; Templeton et al., 1996; Lintsen et al., 2007; Cai et al., 2011; Jones et al., 2011; Nelson and Lawlor, 2011).

Since 1991 all IVF centres in the UK have been obliged to send their data to the Human Fertilisation and Embryology Authority (HFEA), which by law is responsible for the regulation of IVF treatment. The HFEA probably provides the largest, most complete and reliable data set on IVF recordings in the world. The Templeton model (named after the first author) was developed with data on IVF treatment performed between 1991 and 1994 (Templeton et al., 1996), when ICSI was already introduced but hardly used. The Nelson model (named after the first author) was based on IVF treatments performed between 2003 and 2008 (Nelson and Lawlor, 2011) by which time ICSI had become a generally established treatment for male infertility. The Templeton model was often recommended in the past as the best approach, while the Nelson model is characterized by the largest sample size and the most recent approach.

In the present study we compare the performance of both models using independent data, based on a large population of couples with primary infertility treated with IVF in the Netherlands (NL) between 2002 and 2004. We also studied the effect of the changing success rate of IVF over time on the performance of both models.

Patients

The NL IVF cohort 2002–2004

We considered couples who were included in a subsidized national observational cohort study that was carried out in the Netherlands from January 2002 to December 2004 (Lintsen et al., 2007). All IVF centres officially promised to register their IVF treatments. The cohort consisted of 4928 couples with follow-up from the start of IVF or ICSI until the occurrence of an on-going pregnancy or, if no pregnancy arose, until 1 year after the date of the last menstruation prior to the first IVF treatment. If the first IVF cycle did not result in a pregnancy, IVF could be applied in one or more subsequent cycles, and thus in many couples the data from more than one IVF cycle were available.

The following prognostic variables were recorded: maternal age, duration of infertility, cause of infertility, cycle number, primary or a secondary infertility and the type of treatment (IVF or ICSI). The cause of infertility was categorized as unexplained infertility, endometriosis, tubal pathology, cervical-, immunological- or male infertility, which was subdivided in mild, moderate or severe male infertility. Where several causes were possibly involved, one was considered to be the main cause and the reason for infertility in the couple was categorized as such. The term primary infertility was used for couples where the woman had never been pregnant. All female partners of secondary infertile couples had been pregnant previously irrespective of whether the pregnancy resulted in a live birth. All women underwent ovarian stimulation by gonadotrophins and used their own oocytes.

IVF treatment was considered to be successful if an ongoing pregnancy ensued: this was defined as a pregnancy of 8 weeks or more with ultrasound-confirmed fetal heart activity. The original aim of the NL study was to develop a model that predicts the probability of ongoing pregnancy within 12 months after the first IVF cycle. In the present study the NL cohort was used as an independent data
set in which the performance of both UK models was tested (external validation).

The UK models

The Nelson model included the following predictors: maternal age, duration and cause of infertility, the number of treatment attempts in the past, previous pregnancies and live births either naturally conceived or after treatment, source of oocytes (women’s own or donor), type of treatment (IVF or ICSI), hormonal preparations used (gonadotrophins, anti-estrogen or hormone replacement) and cycle number (Supplementary data, A). Cycle number was categorized as 1, 2 or 3 or more. The following causes of infertility were distinguished: unexplained, tubal, anovulation, endometriosis, cervical, male and combination of causes. The predicted outcome was live birth, which was defined as a baby born alive after at least 24 weeks of gestation. The Templeton model included most of the same predictors as the Nelson model with the same categories. Some predictors were not yet considered in the Templeton model, i.e. source of oocytes (only women’s own), type of treatment (only IVF) and hormonal preparations (only gonadotrophins).

Although the variables collected in the UK and NL cohorts were generally similar, some difference exists. Particularly, the UK data contained a detailed subdivision of secondary infertility including the mode of conception (spontaneous or after IVF) and whether there had been a pregnancy or live birth in the past. Since this detailed information was lacking for the NL cohort, we only analysed couples with primary infertility in the comparison. In the Nelson database primary infertility was subdivided into two categories: with and without previous IVF treatment. Since this subdivision has not been used in the NL cohort, we took the weighted averages of the predicted probabilities of these two categories for the validation analysis and model comparison. Further, we used the following categories and adaptations. For the causes of infertility ‘anovulation’ (Nelson) was considered similar to ‘hormonal’ (NL), and ‘cervical’ (Nelson) similar to ‘immunological’ (NL). The outcome of success for both UK models is the occurrence of a live birth after IVF, which is indeed the ultimate outcome of interest. Unfortunately, we could not use live birth because the outcome of success in the NL data is an ongoing pregnancy. In order to make both outcomes compatible, we divided the outcome live birth of both UK models with a factor 0.92 to convert the chance of live birth to ongoing pregnancy. This factor is based on a methodological study by Arce et al. (2005). These authors demonstrate that ongoing pregnancy is an appropriate outcome measure when taking into account that ~8% of the ongoing pregnancies will disappear before reaching the stage of live birth. We choose the overall terms ‘success or no success after IVF’ in the following text to prevent confusion on the outcomes.

Methods of Analysis

Missing values in the NL data ranged from 0% to a maximum of 8% for duration of infertility. The missing data were multiple-imputed (van Buuren, 2012). The models were validated on the 10 completed datasets and analytical results were pooled using Rubin’s rules (Rubin, 1987).

Model performance

We used a corrected version of the Nelson model (Nelson and Lawlor, 2013, see Supplementary data for details) which differed only slightly from the original model (Nelson and Lawlor, 2011). Performance of the previously reported model and updated model was very similar in our data. The performance of the models was evaluated in terms of discrimination and calibration. Discrimination generally measures the ability of a test to distinguish between individuals with and without a given disease. In the context of the present study this refers to the ability to distinguish between women who will or will not have a successful pregnancy after IVF. It can be assessed with the area under the receiver operating characteristic curve (AUC) or c-statistic (Harrell, 1982). Values for the AUC range from 0.5 to 1.0, with higher values indicating better discrimination.

Calibration refers to the level of agreement between the predicted and the observed IVF success rates. It can be assessed graphically with a calibration plot which is defined by its slope and its intercept (Cox, 1958; Harrell, 2001). With perfect calibration, the value of the intercept is equal to 0 and of the slope equal to 1. The intercept measures to what extent the mean predicted success rate equals the overall observed success rate. This concept is referred to as ‘calibration-in-the-large’ and is a measure for the degree of systematic under- or overestimation of the model. Calibration can also be assessed by the predicted-to-observed ratios in arbitrarily chosen equal subgroups with similar predictions for success (e.g. quintiles).

Since the success rates after IVF increased from 14 to 24% over time, (Fig. 1), success rate adjustments were used to improve the calibration-in-the-large of the models (See Supplementary data, B). All analyses were performed using R version 3.0.0 (R Foundation, Vienna, Austria).

Results

Of the 9060 IVF cycles of the NL couples, 63% were from couples with primary infertility. In the NL cohort fewer women were over 35 years of age than in the Nelson and Templeton cohorts (Table I). Moreover, fewer NL women had infertility lasting >3 years. In the Templeton study the cause of infertility was more frequently coded as ‘tubal’. Fewer NL women had three or more cycles that may be explained by the fact that in the NL cohort only cycles performed during the first year following the start of treatment were taken into account. The success rate in the NL data of 21.2% per IVF cycle was closer to the Nelson success rate of 23.4% than to the Templeton rate of 13.9%.

Discrimination and calibration

The discriminative performance of the Nelson and Templeton models on the NL data was similar (Table II). The predicted-to-observed ratios of the original models revealed overestimation of the Nelson model (values above 1) and underestimation of the Templeton model (values smaller than 1). After adjustment for changing success rates, the Nelson model still showed slight overestimation, while the Templeton calibration became almost perfect. These results are further illustrated in Fig. 2. The Nelson model showed systematic overestimation of success rates with line and triangles below the dotted line of perfect calibration (e.g. a predicted probability of 20% corresponds to an observed proportion of ~15%). The Templeton model systematically underestimated the success rates, with the line and triangles above the dotted line (e.g. a predicted probability of 20% corresponds to an observed proportion of nearly 30%). The results indicate that the
adjustments for the success rate of IVF considerably improved the 
calibration of both models, with excellent calibration for the Templeton 
model.

### Table I Description of the predictors as used in the two models for the NL, Nelson and Templeton cohorts

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Category</th>
<th>NL cohort (n = 5176 cycles)</th>
<th>Nelson cohort (n = 163 425 cycles)</th>
<th>Templeton cohort (n = 36 961 cycles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal age, years</td>
<td>18–34</td>
<td>3386 (65.4)</td>
<td>68 008 (41.6)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>35–37</td>
<td>973 (18.8)</td>
<td>40 984 (25.1)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>38–39</td>
<td>500 (9.7)</td>
<td>24 837 (15.2)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>40–42</td>
<td>281 (5.4)</td>
<td>21 218 (13.0)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>43–44</td>
<td>27 (0.5)</td>
<td>5334 (3.3)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>45–50</td>
<td>0 (0.0)</td>
<td>2763 (1.7)</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Missing data</td>
<td>9 (0.0)</td>
<td>281 (0.0)</td>
<td>NA</td>
</tr>
<tr>
<td>Duration of subfertility, years</td>
<td>&lt; 1</td>
<td>103 (2.0)</td>
<td>1799 (1.1)</td>
<td>2258 (6.1)</td>
</tr>
<tr>
<td></td>
<td>1–3</td>
<td>3084 (59.6)</td>
<td>50 278 (30.8)</td>
<td>8407 (22.7)</td>
</tr>
<tr>
<td></td>
<td>4–6</td>
<td>1467 (28.3)</td>
<td>54 738 (33.5)</td>
<td>13 483 (36.5)</td>
</tr>
<tr>
<td></td>
<td>7–9</td>
<td>241 (4.7)</td>
<td>22 173 (13.6)</td>
<td>7017 (19.0)</td>
</tr>
<tr>
<td></td>
<td>10–12</td>
<td>79 (1.5)</td>
<td>9506 (5.8)</td>
<td>3701 (10.0)</td>
</tr>
<tr>
<td></td>
<td>&gt; 12</td>
<td>31 (0.6)</td>
<td>11 219 (6.9)</td>
<td>2092 (5.7)</td>
</tr>
<tr>
<td></td>
<td>Missing data</td>
<td>171 (3.3)</td>
<td>13 712 (8.4)</td>
<td>3 (0.0)</td>
</tr>
<tr>
<td>Cause of infertility</td>
<td>Unknown</td>
<td>1753 (33.9)</td>
<td>44 409 (27.2)</td>
<td>12 340 (31.0)</td>
</tr>
<tr>
<td></td>
<td>Tubal</td>
<td>750 (14.5)</td>
<td>24 734 (15.1)</td>
<td>19 096 (48.0)</td>
</tr>
<tr>
<td></td>
<td>Anovulatory</td>
<td>378 (7.3)</td>
<td>15 304 (9.4)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Endometriosis</td>
<td>502 (9.7)</td>
<td>5463 (3.3)</td>
<td>4117 (10.3)</td>
</tr>
<tr>
<td></td>
<td>Cervical</td>
<td>135 (2.6)</td>
<td>76 (0.0)</td>
<td>4232 (10.6)</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>1397 (27.0)</td>
<td>57 060 (34.9)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Combination causes</td>
<td>–</td>
<td>16 379 (10.0)</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Missing data</td>
<td>261 (5.0)</td>
<td>0 (0.0)</td>
<td>NA</td>
</tr>
<tr>
<td>Type of infertility</td>
<td>Primary</td>
<td>5176 (100.0)</td>
<td>104 967 (64.2)</td>
<td>19 997 (54.1)</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>0 (0.0)</td>
<td>58 458 (35.8)</td>
<td>16 961 (45.9)</td>
</tr>
<tr>
<td></td>
<td>Missing data</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>3 (0.0)</td>
</tr>
<tr>
<td>Cycle number</td>
<td>Cycle 1</td>
<td>2962 (57.2)</td>
<td>93 795 (57.4)</td>
<td>18 239 (49.3)</td>
</tr>
<tr>
<td></td>
<td>Cycle 2</td>
<td>1381 (26.7)</td>
<td>34 860 (21.3)</td>
<td>8123 (22.0)</td>
</tr>
<tr>
<td></td>
<td>Cycle ≥ 3</td>
<td>833 (16.1)</td>
<td>34 770 (21.3)</td>
<td>7339 (19.9)</td>
</tr>
<tr>
<td></td>
<td>Missing data</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>3260 (8.8)</td>
</tr>
<tr>
<td>Treatment Type</td>
<td>IVF</td>
<td>3064 (59.2)</td>
<td>88 244 (54.0)</td>
<td>36 961 (100.0)</td>
</tr>
<tr>
<td></td>
<td>ICSI</td>
<td>2112 (40.8)</td>
<td>75 181 (46.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td></td>
<td>Missing data</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Success after IVF</td>
<td></td>
<td>1097 (21.2)</td>
<td>38 316 (23.4)</td>
<td>5138 (13.9)</td>
</tr>
</tbody>
</table>

Values are numbers (%). 
NA, numbers were not/partly available. 
*aCategories differed for the three cohorts. See text.

### Table II Discrimination and calibration of the Nelson and Templeton models in the NL cohort

<table>
<thead>
<tr>
<th></th>
<th>Nelson model</th>
<th>Templeton model</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.590</td>
<td>0.597</td>
</tr>
<tr>
<td>Predicted-to-observed ratios of the original models</td>
<td>1.36 (1.23–1.45)</td>
<td>0.68 (0.64–0.76)</td>
</tr>
<tr>
<td>Predicted-to-observed ratios of the adjusted models</td>
<td>1.15 (1.00–1.21)</td>
<td>0.94 (0.90–1.05)</td>
</tr>
</tbody>
</table>

AUC, area under the receiver operating characteristic curve. 
*aOverall ratio with range of five quintiles between brackets. 
*bAdjusted for overall IVF success rate.

### Discussion

In this study, we compared the performance of two UK models to predict success after IVF treatment in a Dutch cohort of couples with primary infertility. The Templeton and Nelson models showed similar ability to distinguish between success and no success after IVF. The Templeton model systematically underestimated the chance of success in the NL cohort while the Nelson model overestimated the chance of success. Since IVF success rates have almost doubled since the beginning of the 1990s, the underestimation of the Templeton model is not surprising and the calibration greatly improved after taking the time trend into account. The Nelson model also performed much better after adjustment for success rates.

Although discrimination and calibration are the accepted measures to assess the quality of prediction models, there is an ongoing debate as to whether both are equally important with regard to predicting a pregnancy leading to a live birth occurring spontaneously or after IVF (Cook, 2008; Pencina et al., 2008; Coppus et al., 2009; Leushuis et al., 2009).
Discrimination assesses to what extent a model is able to distinguish couples with a successful treatment from unsuccessful couples. By analogy with diagnostic problems which deal with two populations of diseased and non-diseased individuals, discrimination applied in prediction models implicitly assumes that the population of couples with a fertility problem consists of two subpopulations: couples who have not yet conceived but will do so within the foreseeable future and sterile couples who will never conceive. However, the fertility potential of couples presenting for IVF varies from complete sterility to normal fertility. In the absence of strong predictors, good discrimination between sterile and fertile couples is difficult and it has been suggested that the maximum AUC value that can be reached in this situation is $\approx 0.62$ (Coppus et al., 2009).

Calibration evaluates to what extent the predicted and the observed success rates are in agreement. This approach is consistent with the current concept of subfertility care in which treatment should only be considered when the chance of success after treatment exceeds the spontaneous likelihood of it occurring. As a consequence, IVF prediction models with a relatively low discriminatory ability can still be useful if predicted probabilities of success cover a clinically relevant range which does not (or partly) overlap the predicted probabilities of spontaneous conception.

External validation in independent data, for example data from other countries, is important to assess model performance. Cultural and national differences between countries can influence the result of external validation. However, the medical cultures of the UK and Netherlands seem rather alike. For example the National Institute for health and Care Excellence guidelines (UK) and the Dutch Society of Obstetrics and Gynaecology (NVOG) guidelines (NL) describe similar approaches for infertility management. Further, the Dutch validation population of the present study was also used to develop a prediction model for IVF treatment success. The same anamnestic variables as in the Templeton and Nelson models were identified as important predictors (Lintsen et al., 2007).

Why does the Nelson model not perform better than the older Templeton model? We consider two possible reasons. First, the impact of the leading predictive variables (maternal age, the duration of the infertile period, cycle number) is likely to be stable over time resulting in similar discriminatory power of the two models. Indeed, the discrimination found in the NL cohort was similar for both models. Yet, compared...
with the Templeton model, the Nelson model contains two new, potentially promising predictors: source of oocyte (own versus donor) and type of treatment (IVF versus ICSI). Source of oocyte is a strong predictor (multivariable odds ratio (OR) = 2.63 for donor versus own oocyte) but the rare use of donor oocytes (4% in the Nelson cohort and 0% in NL cohort) implies that the effect on the performance of the model is low. In contrast, the use of ICSI is much more frequent (46% in the Nelson cohort and 41% in the NL cohort); however, it has a weaker predictive effect (multivariable OR = 1.27). Apparently, inclusion of these new predictors did not improve the performance of the Nelson model compared with the Templeton model.

Second, the Nelson model is possibly too closely fitted to the data, which is reflected by irregularities in predicted chances that may arise. For example, according to the online calculator developed from the model, a 35-year-old woman who uses her own oocytes and has been trying to become pregnant for 5 years has a higher chance of success after IVF compared with a similar woman who has been trying for 2 years. Such biologically implausible outcomes may occur in small subgroups that are the result of categorizing continuous variables. In the Nelson model the predictors ‘age’ and ‘duration’ are categorized in many groups and the age effects vary by duration (interaction). As a consequence, the Nelson model predicts well in the population on which it is fitted but might fail when applied in new data, despite the large size of the Nelson cohort, which usually prevents such over fitting (Harrell, 2001; Steyerberg, 2008).

This stresses the importance of testing models in populations other than the one in which it was developed. Good performance at external validation is considered a requirement before the introduction of a model in clinical practice (Altman and Royston, 2000; Reilly and Evans, 2006; Steyerberg, 2008; Leushuis et al., 2009; Moons et al., 2009). In our case, we found a discrepancy between the performance of the Nelson model in the data from which it was developed and the performance in the independent NL cohort. Discrimination was 0.633 in the data from which it was developed and 0.590 in the NL data, while calibration was excellent in the data from which it was developed but showed some signs of overestimation in the NL data. Such a difference can be the result of either different predictors or a different case mix (Vergouwe et al., 2010).

Another known and very strong predictor which has not been included in both UK models is the treatment centre (Lintsen et al., 2010). Using a random effects model to include all centres is an efficient way to estimate the effects of other predictors (e.g. age) conditional on the centres. However, since both models were applied in centres that were not included in the data used to develop the model, the centre effect in new patients is unknown and only the fixed effects of the predictors can be used to calculate IVF success. Moreover, the predictor effects that are corrected for the centre effects may be similar to the unadjusted effects (Bouwmeester et al., 2013) and taking centre into account will then not much change the predictions. Nevertheless, it would be worthwhile to study the effect of centre adjustment in future prediction models for IVF treatment success. The Templeton model was tested previously in a 1991 - 1999 NL population (Smeenk et al., 2000) and more recently in a 2001 - 2009 NL population (van Loendersloot et al., 2011). In a review on prediction models in reproductive medicine, the authors concluded that the Templeton model is the only externally validated model reliably predicting prognostic groups after IVF (Leushuis et al., 2009). However, as we demonstrated here, this conclusion is only valid when the time trend of improved success rates is taken into account. Calibration of the Templeton model was poor in the present NL population but considerably improved after adjusting for better success rates. In a recently published French study the Templeton model was validated in a single centre (Arvis et al., 2012). Also in this study calibration of the original model was poor but could easily be adjusted for time trend. Apparently, the Templeton model is able to capture the variability due to patient mix and its impact on predicted success categories.

Only few effective predictors could be selected from the HFEA database and adding other variables may improve the performance of the model. Indeed, Arvis and co-workers added the basal FSH level, the BMI and the smoking habits of their patients to the predictors of the Templeton model and found a considerable improvement in performance (Arvis et al., 2012). However, the value of this addition has still to be validated in an independent data set. Other potential predictors not yet evaluated are the level of anti-Mullerian hormone, the antral follicle count or the age of the male partner. When the actual treatment has already been started, the number of oocytes obtained, their fertilization rate, the quality of the embryos and, especially, the number of replaced embryos have been shown to considerably improve the prediction of live birth (Cai et al., 2011; Jones et al., 2011). However, this information is not yet available for couples who have to be counselled before treatment when it is most needed.

For several years, frozen embryo replacement (FER) has become a successful trend in IVF practice. Embryos that are not used for replacement during the cycle in which the oocytes are retrieved are cryopreserved. The embryos are then replaced months or even years later during a normal menstrual cycle; the success rate per cryopreserved embryo replacement appears to be at least as good as fresh embryo replacement. When single embryo transfer is used—strongly recommended by all stakeholders involved in IVF—this approach offers great advantages, because hardly any multiple pregnancies occur. Multiple pregnancies are considered to be a serious complication of conventional IVF because of the high chance of premature births associated with increased risks of infant mortality, morbidity and cognitive and neurological problems later in life (Helmerhorst et al., 2004). If live birth ensues after transfer of these so-called rest embryos, the success is to be attributed to the IVF cycle during which the oocytes were obtained. The recent European Society of Human Reproduction and Embryology report on the 2008 IVF results in Europe demonstrates that in the UK live birth rates per cycle of ~25% per fresh cycle rose to almost 30% when the results of FER were included (Ferraretti et al., 2012). The increased use of FER will require further adjustments of the prediction models to provide accurate estimates of success rate changes over time.

Our study has several limitations. First, we could validate only the models in couples with primary infertility because detailed information on secondary infertile couples was lacking in the NL test population which was therefore relatively homogeneous. This may have influenced the performance of particularly the Nelson model (Vergouwe et al., 2010). Nevertheless, our study shows that reliable predictions can be given in the primary infertility group. Apart from being the largest (about two thirds of the target population), it is, in our opinion, also the most relevant group because secondary infertile couples have proved their fertility while many of them have at least one child. Second, our validation population consisted only of couples that used their own oocytes. The predictor ‘donor oocytes versus own oocytes’ in the Nelson model was useless when validating the model because no patients who used donor oocytes were included in the NL data set.
Since donor oocytes were only used by a small fraction of patients in the Nelson data set, it is unlikely it would have a large impact on the performance of the Nelson model.

In conclusion, the more recently developed Nelson model did not perform better than the older Templeton model in predicting IVF success rates in Dutch couples with primary infertility. We found that adjustment of the models was required to account for changes over time in IVF success rates.

**Supplementary data**

Supplementary data are available at http://humrep.oxfordjournals.org/.

**Authors’ roles**

E.R.t.V. developed the concept of the study and wrote the first draft; Y.V. had a major contribution in the Methods of Analysis and Results, and Discussion sections, Y.V., E.R.t.V., J.D.F.H. and D.N. contributed to the design of the study; D.N. performed the statistical analyses; all authors contributed intellectually to the final version of the manuscript.

**Funding**

No external funding was sought or obtained.

**Conflict of interest**

None declared.

**References**


