Simple adaptations to the Templeton model for IVF outcome prediction make it current and clinically useful

P. Arvis¹,*, P. Lehert²,³, and A. Guivarc’h-Levêque¹

¹Clinique la Sagesse, Place St Guénolé, Rennes 35000, France ²Statistics Department, Faculty of Economics, University of Louvain, UCL Mons, 151 Chaussée de Binche, 7000 Mons, Belgium ³Faculty of Medicine, The University of Melbourne, Melbourne, Australia

*Correspondence address. 430 route de Pessicart 06100 Nice, France, Tel: +33-492-09-82-42; E-mail: philippe.lehert@gmail.com

Submitted on December 11, 2011; resubmitted on June 18, 2012; accepted on June 26, 2012

STUDY QUESTION: What is the validity of the Templeton model (TM) in predicting live birth (LB) for a couple starting an IVF/ICSI cycle?

SUMMARY ANSWER: A centre-specific model based on the original predictors of the TM may reach a sufficient level of accuracy to be used in everyday practice, with a few simple adaptations.

WHAT IS KNOWN AND WHAT THIS PAPER ADDS: The TM seems the best predictive model of LB in IVF. However, previous validations of the TM suggest a lack of discrimination and calibration which means that it is not used in regular practice. We confirm this finding, and argue that such results are predictable, and essentially due to a strong centre effect. We provide evidence that the TM constitutes a useful reference reflecting a high proportion of the patient-mix effect since the parameters of the model remain invariant among centres, but also across various cultures, countries and types of hospitals. The only difference was the intercept value, interpreted as the measurement of the global performance of one centre, in particular, for a population of reference.

STUDY DESIGN: The validity of the TM was tested by a retrospective analysis of all IVF/ICSI cycles (n = 12,901) in our centre since 2000.

PARTICIPANTS, SETTING AND METHODS: All IVF/ICSI cycles were included in the analysis. The model discrimination was evaluated by C-statistics, calculated as the area under the curve of an ROC curve. The TM was then adjusted for our data and additional variables were assessed.

MAIN RESULTS AND THE ROLE OF CHANCE: Poor calibration and discrimination (C = 0.64) was observed in conformity with previous external validations. Fitting the TM to our centre constituted the first substantial improvement in prediction accuracy of discrimination (C = 0.69) and calibration. We identified an important linear time trend effect and the added value of three other predictors (FSH, smoking habits and BMI) that significantly improved the model (C = 0.71).

BIAS, CONFounding AND OTHER REASONS FOR CAUTION: Bias due to missing data handling was assessed through sensitivity analyses.

GENERALIZABILITY TO OTHER POPULATIONS: Neither the TM nor any other models based on some centres are directly applicable to other centres. However, the TM constitutes a useful basis to build an accurate centre-specific model.

STUDY FUNDING/COMPETING INTEREST(S): There were no commercial relationships (i.e. consultancies, patent-licensing agreements) that might pose a conflict of interest in connection with the submitted manuscript. The objective of this research was not directed toward any treatment effects.

Key words: predictive model / live birth / IVF / ICSI

Introduction

Determining which baseline characteristics are associated with the highest chance of achieving a successful pregnancy and live birth (LB) with IVF has received attention in recent assisted reproduction technology (ART) literature. IVF is costly and not exempt of risk. Thus to evaluate a priori the chances of success for a couple in using IVF based on baseline conditions may constitute a key decision-making tool for both the physician and patients. Our main motivation in starting this project was to set up a predictive model providing the best
possible prediction of LB when a couple decides to start a new IVF/ICSI cycle.

The development of a predictive model of LB has received interest in recent literature, with numerous approaches (limited to IVF data): Templeton et al. (1996, n = 36,961), Commenges-Ducos et al. (1998, n = 923 cycles), Minaretzis et al. (1998, n = 544), Stolwijk et al. (1996, n = 1,424), Bancsi et al. (2000, n = 435 patients), Stolwijk et al. (2000, n = 1,315), Smeenk et al. (2000, n = 1,253), Hunault et al. (2002, n = 642), Ferlitsch et al. (2004, n = 170), Ottosen et al. (2007, n = 2,193) and Lintsen et al. (2007, n = 4,928). However, only a few have received appropriate validation.

The Templeton model (TM) was the most frequently used and referenced, and considered as the best choice in a recent review (Leushuis et al., 2009). Importantly for our project, it is only constituted by patient-oriented variables or patient-mix (socio-demographic or fertility history), excluding information only available later in the cycle (number of oocytes, embryos, etc.). A retrospective validation by Smeenk et al. (2000), confirmed the predictive ability of the TM to reasonably discriminate between low and high prognoses, but also confirmed the poor predictive performances of the model. A more recent validation (Van Loendersloot et al., 2011) similarly confirmed poor discrimination and inadequate calibration, and a systematic underestimation of LB. Furthermore, predictive models such as the TM were developed almost exclusively in the few countries where data collections were made available from early individual recording of IVF cycles as requested by national health authorities (UK and Netherlands).

The lack of prediction accuracy, and the extent to which country-specific variables such as lifestyle and socio-cultural differences interfere with pregnancy chances, in addition to clinical variables, required assessment of the validity of the TM before routine use; this constituted our first and main objective. When this analysis provided evidence of insufficient accuracy, supplementary considerations of model development were planned, based on the following rationale: (i) The TM needed to be developed to predict LB irrespective of centre-specific conditions. The high heterogeneity between centres constitutes one of the major sources of variability in ART performance (Lintsen et al., 2010), and is mainly due to differences in the procedures used, specific treatments, know-how, biological protocols, etc. Thus, the predictive performance of a model exclusively based on patient-mix is likely to improve if we adapt the model to one centre in particular, by fitting the model to its own data. (ii) The original TM could be out-dated (van Loendersloot et al., 2011), as developments in ART have significantly improved LB rates, a factor which may explain the low predictive power of the model. For example, the ICSI technique was unknown when the model was constructed. (iii) As it was developed on the basis of a very large sample size [Human Fertilization and Embryology Authority (HFEA)], the TM was limited to the available covariates. Alternative models have been proposed but based on much smaller samples: ovarian reserve estimates (basal FSH or FSH/LH ratios) (Sharif et al., 1998), Day 3 estradiol (Smotrich et al., 1995), anti-Müllerian hormone (AMH) levels, oocyte counts or embryo quality (Craft and Forman, 1997), among others. Therefore, we questioned the interest of new predictors likely to increase the accuracy of the TM when added to the original model.

### Materials and Methods

#### Data collection

All IUI/IVF/ICSI cycles performed in our centre from 2000 onwards were recorded through a Data Entry system. There were 12,500 IVF/ICSI cycles available for this study. For each cycle, a set of baseline characteristics, stimulation parameters and pregnancy outcomes were recorded. A retrospective analysis was conducted based on all documented IVF/ICSI cycles, including those with early interruption at any stage. The available variables were: female age, male age, duration of infertility (DIFF), pregnancy history (primary/secondary infertility), cause of infertility (tubal infertility (TBI), male infertility, cervical infertility, ovastral dysfunctions, endometriosis, unexplained infertility), uterus abnormality, weight, BMI (weight/height²), basal FSH (IU/L), number of previous failed cycles (NFC), smoking habits of women and partners (binary, Yes/No), previous number of miscarriages, DIFF, previous IVF pregnancies resulting in LB (LB), previous IVF pregnancies not resulting in LB (INB), previous non-IVF pregnancies resulting in LB (NLB), previous non-IVF pregnancies not resulting in LB (NNB). Antral follicle count (AFC) and AMH (ng/ml) data were only available since 2008. Data for stimulation outcomes (during down-regulation, stimulation and post-triggering period) were not used in this analysis, as the model should predict LB before the decision to start the cycle is taken.

#### Statistical analysis

As LB was our main end-point, we limited our study to cycles starting no later than July 2011 in order to include cycles with known outcome. The predicted probability (P) of achieving a pregnancy after IVF was calculated using the TM: 

\[
P = \frac{1}{1 + e^{-\left(\logit L - 2.028 + 0.00551 \times (age - 16)^2 - 0.00028 \times (age - 16)^3 + DIF - 0.0690 \times NFC - 0.0711 \times TBl + 0.7587 \times ILB + 0.2986 \times INB + 0.2277 \times NLB + 0.1117 \times NN_B\right)}}
\]

The DIF values were 0.2163, −0.089, −0.1036, −0.4179 based on whether the duration of infertility was ≤3 years, 4–6 years, 7–12 or ≥13 years, respectively.

We first validated the TM. The model discrimination (extent to which the model allows discrimination of successes from failures) was evaluated by C-statistics calculated as the area under the curve (AUC) of an ROC curve. A model was considered to have a poor, fair or good performance if the AUC ranged between 0.5 and 0.7, 0.7 and 0.8 or 0.8 and 0.9, respectively (Swets, 1988). Model calibration is known to be at least as important as discrimination in the context of IVF (Coppus et al., 2009), and was evaluated by the goodness-of-fit test (Hosmer, 2000) with a correction factor (Harrell et al. 1996), plotting the observed and predicted frequencies for 10 equally sampled categories, and comparing the regression line with the diagonal (perfect calibration).

At a second stage, we assessed the extent to which the TM might be improved. First we re-fitted the TM model on our own data (Model I). We used a bootstrapping technique for estimation of confidence intervals (CI), and a shrinkage factor to reduce the overfit of the model and to obtain relatively unbiased estimates (Steyerberg et al., 2001). For each parameter estimate, we tested the significance of the difference between the TM and our own estimate.

At a later stage, we tested the significance of trend in time (LRT likelihood ratio test) considered both as a main effect and as an interaction with the other variables. Finally, we assessed the added value of potential predictors not included in the original TM, but suspected to be an additional effect in previous studies (Model II).

Starting from the original TM, we tested the significance of adding a new variable (hierarchical test on the nested model) to the increase of likelihood ratio, by selecting predictors providing a significant increase of the Akaike Information Criterion. We also examined the net benefit of
development Core Team (2008). Net reclassification improvement (NRI) constitutes the net effect on reclassification tables constructed separately for participants with and without events, and quantifies the correct movement in categories upwards for events and downwards for non-events. The integrated discrimination improvement (IDI) focuses on differences between sensitivity and specificity for models with and without the new predictors. Continuous variables were categorized (generalized additive models, Hastie and Tibshirani, 1987). The considered cutting-points to categorize these variables were not selected through the analysis, but were chosen from literature findings: smoking habits had only two categories (never smoked versus past or current smoker). For a BMI, from the available continuous value (weight/height$^2$), we fixed the categories following WHO recommendations in considering normo-weighted women as $18 < \text{BMI} < 26$. For FSH, a cut-off value of $\text{FSH} = 10$ IU/L was selected in compliance with the studies of Abdalla and Thum (2004) and Chuang et al. (2003).

To account for more than one cycle for the same patient, the predictive model was fitted by using a non-linear mixed model featuring a logistic model where the patient was considered as a random factor. A simple logistic linear model was conducted for sensitivity purposes. Given the large size of the data base, to guarantee that statistical significance was associated with clinical relevance, the main effects were tested at 0.1% confidence level, and interactions at 5%. All results were reported with 95% CI. As some variables included in this validation study were not available in the data base (i.e. ILB, INB, NLB, NNB) additional information was searched (in archives or if necessary from the patient). A mixed model was used to implement the Full Information Maximum Likelihood technique, for handling missing data. We checked the sensitivity of the model to missing data imputation techniques by comparing results from the whole population, pairwise deletion (any pair of non-missing values used in the Covariance matrix) and listwise deletion consisting of selecting patients with no missing data. Sensitivity to missing data was estimated by calculating the mean coefficient of variation on three missing data imputation techniques over all the coefficients of the model. All our statistical analyses were carried out with the help of R statistical package (R, version 2.12.2) R Development Core Team (2008).

## Results

### Sample description

A total of 12 901 cycles were available from January 2000 to June 2011 (last observations were not used as LB was not yet known) from consecutive collection of all cycles produced in the clinic, including all premature interruptions (intent-to-treat selection) and which constituted our main data sample with the following general characteristics (Table I). The median age was 33 years (inter quartile range: 30–36). The median number of failed cycles was two with extreme maximum values of 13. Clinical pregnancy rate was 19.2% and LB rate was 17.8%. Missing data occurred in the data base with the following frequency: duration of infertility (1.5%), diagnostic category (6.1%), pregnancy history (4.1%), smoking habits (18.9%) and weight (22.1%). AMH was only available for recent cycles ($n = 1569$). For 1693 patients, despite all efforts to find missing information, at least one variable required for validation of the TM was not available, thus 11 208 cycles were fully documented on a per protocol basis.

### TM validity

An observed discrimination of $C = 0.64$ (95% CI: 0.62–0.65) was found. As suggested by Smeenk et al. (2000), we repeated this calculation using women’s age at the first IVF cycle, and not at the current cycle, which resulted in a virtually unchanged value. We tested model calibration by comparing mean expected and observed LB estimates within ten categories determined by the observed deciles of the distribution of the predicted value. The difference between the observed and predicted pregnancy rates varied from $-2.35$ to $9.50\%$ in particular for large values (Hosmer test, $P < 0.001$, Fig. 1).

## Centre-specific fitting based on the same variables

We re-fitted the TM model based on exactly the same variables, with a bootstrapping technique for estimation of CIs, and a shrinkage factor (Table II). The mixed model accounting for several cycles for the same patient and simple logistic regression provided almost the same results.

We observed a highly significant improvement in discrimination ($C = 0.69$; 95% CI: 0.67–0.70), and almost perfect calibration (Fig. 1), with $<1.5\%$ difference between the true LB value and the prediction in each decile subgroup and the fitted line coinciding with the diagonal (slope = 0.96; 95% CI: 0.88–1.01).

We tested the significance of the difference between the original TM and our own estimate for each parameter. No difference reached the significance threshold ($P > 0.05$), except for the intercept.

### Table I: Sample description.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$n$</th>
<th>95% CI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>33$^1$</td>
<td>30–36$^2$</td>
</tr>
<tr>
<td>Age of partner</td>
<td>35$^1$</td>
<td>31–39$^3$</td>
</tr>
<tr>
<td>Infertility duration</td>
<td>5.45</td>
<td>± 2.60</td>
</tr>
<tr>
<td>BMI</td>
<td>22.55</td>
<td>± 4.54</td>
</tr>
<tr>
<td>BMI of partner</td>
<td>24.70</td>
<td>± 3.33</td>
</tr>
<tr>
<td>Number of failing cycles</td>
<td>$^2$2$^1$</td>
<td>$^2$1–3</td>
</tr>
<tr>
<td>Baseline FSH (IU/L)</td>
<td>7.45</td>
<td>± 5.10</td>
</tr>
<tr>
<td>Baseline AMH (IU/L)</td>
<td>3.38</td>
<td>± 4</td>
</tr>
<tr>
<td>Ratio FSH/LH</td>
<td>1.81</td>
<td>± 1.35</td>
</tr>
<tr>
<td>Previous LB with IVF</td>
<td>(1050)</td>
<td>8.1%</td>
</tr>
<tr>
<td>Previous non-LB with IVF</td>
<td>(1895)</td>
<td>14.7%</td>
</tr>
<tr>
<td>Previous LB not by IVF</td>
<td>(1980)</td>
<td>15.3%</td>
</tr>
<tr>
<td>Previous non-LB not by IVF</td>
<td>(1844)</td>
<td>14.3%</td>
</tr>
<tr>
<td>Smoking habits</td>
<td>(2432)</td>
<td>32.4%</td>
</tr>
<tr>
<td>Smoking habits of partner</td>
<td>(2302)</td>
<td>43.6%</td>
</tr>
<tr>
<td>Infertility cause</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infertility male</td>
<td>(3264)</td>
<td>48.0%</td>
</tr>
<tr>
<td>Tubal pathology</td>
<td>(1085)</td>
<td>16.0%</td>
</tr>
<tr>
<td>Endometriosis</td>
<td>(693)</td>
<td>10.2%</td>
</tr>
<tr>
<td>Unexplained</td>
<td>(1039)</td>
<td>15.3%</td>
</tr>
<tr>
<td>Ovulatory problems</td>
<td>(718)</td>
<td>10.6%</td>
</tr>
<tr>
<td>Uterus abnormality</td>
<td>(516)</td>
<td>4.6%</td>
</tr>
<tr>
<td>Clinical pregnancy</td>
<td>(2477)</td>
<td>19.2%</td>
</tr>
<tr>
<td>LB</td>
<td>(2291)</td>
<td>17.8%</td>
</tr>
</tbody>
</table>

$n = 12\,901$ cycles. Mean ± SD or (counts) and percentage (%), except for age and number of failing cycles reported as median (1) and interquartile range (2).
for which values of $-2.03$ and $-1.07$ were found for the TM and our model, respectively ($P < 0.0001$, Table II).

Testing a longitudinal trend in time

Further to the recent hypothesis of increasing bias of the TM with time, we assessed the potential change in accuracy of the TM by separately calculating $C$-statistics for the four 5-year categories from 2000. $C$-statistics were 0.68, 0.67, 0.67 and 0.66 for these consecutive periods. We re-calculated our Model I using the year of the cycle (year 2011 considered as the reference) as a surrogate continuous variable for which values of $-2.03$ and $-1.07$ were found for the TM and our model, respectively ($P < 0.0001$, Table II).

Incorporating new predictors

We tested the added value of three potential predictors already mentioned in previous studies and available in our database: Basal FSH (IU/L), BMI and smoking habits. All the FSH values were available, although BMI and smoking habits were missing in 22.1 and 18.9% of the whole files, respectively. Other variables expected to increase discrimination (AFC and AMH) were only reported for the most recent cycles (n < 2008), and thus were not incorporated into this analysis. An FSH value lower than 10 IU/l was found to be a positive binary predictor (OR = 2.41; 95% CI: 1.98–2.92; $P < 0.001$). We considered normo- and overweighted women as defined by a BMI such that $18 <$ BMI $< 26$ (WHO guidelines). An increased LB was observed for normo- and overweighted patients (OR = 1.29; 95% CI: 1.14–1.45). Finally, women reported as past/current smokers were associated with a negative prognosis (OR = 0.60; 95% CI: 0.52–0.68). Our final Model II incorporating these three variables was characterized by $C = 0.71$, (95% CI: 0.68–0.74), and calibration was identical to our initial Model I.

We compared the classification of subjects with and without LB events to assess improvement with new predictors (Model II) compared with the first model. For 460 subjects with final LB classification, improved using Model II, and for 260, it became worse, with a net gain in reclassification proportion of 0.171, significantly greater than zero ($P < 0.001$). The net gain in reclassification proportion for subjects who did not achieve LB ($n = 7044$) was not significant; 1709 women were reclassified downwards and 1661 were reclassified upwards ($P = 0.047$). NRI was estimated at 0.178 and was highly significant ($P < 0.001$). IDI was estimated at 0.016 ($P < 0.001$). Thus the addition of FSH, smoking habits and BMI improved classification for a net 17.8% of women with events, without net loss for non-events.

Discussion

Limitations

The Templeton study was based on a very large sample size ($n = 36,961$). Data were collected from the HFEA data base and issued from many centres. Submitted to regulatory requirements, the data are presumed of good quality. Our sample was much smaller ($n = 12,901$), although it probably constitutes the second largest data base used for validating the TM. Missing data were inevitable, in particular, for new variables. As our main motivation in this research was to predict the probability of success before starting the cycle (thus solely based on patient-mix and not taking stimulation outcomes into consideration), we disregarded treatment characteristics (down-regulation, stimulation and post-triggering), despite the expected high determination of these variables (Van Loendersloot et al., 2010).

TM validation

The poor level of discrimination ($C = 0.64$) is comparable with previous validations: $C = 0.63$ (Templeton et al., 1996), 0.63 (Smeenk et al., 2000) and 0.61 (Van Loendersloot et al., 2011). Most were based on single academic centre data (Nijmegen, the Netherlands, $n = 1292$, and Amsterdam, the Netherlands, $n = 1394$), with the latter based on ongoing pregnancy and not on LB. Compared with these studies, the present study is characterized by a larger sample size, spread over a longer period of time (2000–2011) and based exclusively on the LB rate.

Calibration was similarly characterized by a poor performance: as concluded in the most recent validation (Van Loendersloot et al., 2011), but in contrast to earlier validations (Smeenk et al., 2000; Leushuis et al., 2009), we found a strong under-estimation of prediction for almost all levels of predictions, but a particularly marked difference in categories with high values (Fig. 1) with estimation errors as large as 9.5%.

The TM is based exclusively on patient-mix variables, although the variability explained by patient-mix has been found to be much lower than the centre effect, considered the most important predictor (Lintsen et al., 2010). Using the TM fitted from a group of referent centres into another study centre will inevitably exhibit a systematic under- or over-estimation depending on whether the performance

Figure 1 Calibration plot. A comparison of the mean probability in each decile subgroup versus the observed proportion was plotted. Points corresponding to our Model I are represented by (+), the vertical line featuring the 95% CI. Points corresponding to the original TM are represented by (open triangle). In case of perfect calibration, all points in the calibration plot should be located on the diagonal, the line of equality.

for which values of $-2.03$ and $-1.07$ were found for the TM and our model, respectively ($P < 0.0001$, Table II).

Testing a longitudinal trend in time

Further to the recent hypothesis of increasing bias of the TM with time, we assessed the potential change in accuracy of the TM by separately calculating $C$-statistics for the four 5-year categories from 2000. $C$-statistics were 0.68, 0.67, 0.67 and 0.66 for these consecutive periods. We re-calculated our Model I using the year of the cycle (year 2011 considered as the reference) as a surrogate continuous variable for which a significant effect was found [odds ratio (OR) = 0.96, (95% CI: 0.94–0.98), $P < 0.001$]. We failed to find any significant interaction between time and all the other predictors ($P > 0.05$).

Incorporating new predictors

We tested the added value of three potential predictors already mentioned in previous studies and available in our database: Basal FSH (IU/L), BMI and smoking habits. All the FSH values were available, although BMI and smoking habits were missing in 22.1 and 18.9% of the whole files, respectively. Other variables expected to increase discrimination (AFC and AMH) were only reported for the most recent cycles (n < 2008), and thus were not incorporated into this analysis. An FSH value lower than 10 IU/l was found to be a positive binary predictor (OR = 2.41; 95% CI: 1.98–2.92; $P < 0.001$). We considered normo- and overweighted women as defined by a BMI such that $18 <$ BMI $< 26$ (WHO guidelines). An increased LB was observed for normo- and overweighted patients (OR = 1.29; 95% CI: 1.14–1.45). Finally, women reported as past/current smokers were associated with a negative prognosis (OR = 0.60; 95% CI: 0.52–0.68). Our final Model II incorporating these three variables was characterized by $C = 0.71$, (95% CI: 0.68–0.74), and calibration was identical to our initial Model I.

We compared the classification of subjects with and without LB events to assess improvement with new predictors (Model II) compared with the first model. For 460 subjects with final LB classification, improved using Model II, and for 260, it became worse, with a net gain in reclassification proportion of 0.171, significantly greater than zero ($P < 0.001$). The net gain in reclassification proportion for subjects who did not achieve LB ($n = 7044$) was not significant; 1709 women were reclassified downwards and 1661 were reclassified upwards ($P = 0.047$). NRI was estimated at 0.178 and was highly significant ($P < 0.001$). IDI was estimated at 0.016 ($P < 0.001$). Thus the addition of FSH, smoking habits and BMI improved classification for a net 17.8% of women with events, without net loss for non-events.

Discussion

Limitations

The Templeton study was based on a very large sample size ($n = 36,961$). Data were collected from the HFEA data base and issued from many centres. Submitted to regulatory requirements, the data are presumed of good quality. Our sample was much smaller ($n = 12,901$), although it probably constitutes the second largest data base used for validating the TM. Missing data were inevitable, in particular, for new variables. As our main motivation in this research was to predict the probability of success before starting the cycle (thus solely based on patient-mix and not taking stimulation outcomes into consideration), we disregarded treatment characteristics (down-regulation, stimulation and post-triggering), despite the expected high determination of these variables (Van Loendersloot et al., 2010).

TM validation

The poor level of discrimination ($C = 0.64$) is comparable with previous validations: $C = 0.63$ (Templeton et al., 1996), 0.63 (Smeenk et al., 2000) and 0.61 (Van Loendersloot et al., 2011). Most were based on single academic centre data (Nijmegen, the Netherlands, $n = 1292$, and Amsterdam, the Netherlands, $n = 1394$), with the latter based on ongoing pregnancy and not on LB. Compared with these studies, the present study is characterized by a larger sample size, spread over a longer period of time (2000–2011) and based exclusively on the LB rate.

Calibration was similarly characterized by a poor performance: as concluded in the most recent validation (Van Loendersloot et al., 2011), but in contrast to earlier validations (Smeenk et al., 2000; Leushuis et al., 2009), we found a strong under-estimation of prediction for almost all levels of predictions, but a particularly marked difference in categories with high values (Fig. 1) with estimation errors as large as 9.5%.

The TM is based exclusively on patient-mix variables, although the variability explained by patient-mix has been found to be much lower than the centre effect, considered the most important predictor (Lintsen et al., 2010). Using the TM fitted from a group of referent centres into another study centre will inevitably exhibit a systematic under- or over-estimation depending on whether the performance
the variability due to the patient-mix to predict LB. Irrespective of or types of hospital. In a Latin country and a non-academic centre, our study suggests that except values, where a huge difference was found. Moreover, conducted to the corresponding estimates in the original TM except for the inter-

an almost perfect calibration was found. Fig. 1. The estimates of each parameter calculated on our data were not significantly different to the corresponding estimates in the original TM except for the intercept values, where a huge difference was found. Moreover, conducted in a Latin country and a non-academic centre, our study suggests that the TM remains applicable across different medical cultures, countries or types of hospital.

Centre-specific fitting

Our Model I constitutes a simple re-fitting of the TM based on the same variables. Discrimination was much improved (C = 0.69), and an almost perfect calibration was found (Fig. 1). The estimates of each parameter calculated on our data were not significantly different to the corresponding estimates in the original TM except for the intercept values, where a huge difference was found. Moreover, conducted in a Latin country and a non-academic centre, our study suggests that the TM remains applicable across different medical cultures, countries or types of hospital.

These results provide some evidence of the ability of TM to capture the variability due to the patient-mix to predict LB. Irrespective of centre differences, a similar effect of each predictor was found; however, the difference between the intercepts can be interpreted in terms of difference in centre performance. The intercept provides an estimate of the LB rate for a reference population depending on how the predictors were coded. When all the predictors are fixed at zero, the intercept determines the probability of LB for this reference population (The logistic model limited to the intercept is 1/(1 + exp(-intercept))). For instance for age = 16 years, an estimate of −1.07 corresponds to a LB rate of 25%.

Model I is TM adjusted for our centre. Model II incorporates all the new variables including year trend effect, FSH (binary FSH > 10), BMI (patient such that BMI < 18 or BMI > 26) and smoking habits (patient smokes or smoked in the past).
with the reference group. The original TM fitting considered the reference age of women as 16 years. We first used the same definition to compare results, but 16 years is uncommon for IVF, and an alternative reference population would be better: 30 years is very close to the mean age of our sample and is close to the optimal maternal age in terms of fertility, thus this definition of the reference population corresponds to a frequently observed population defined as a 30-year-old woman, undergoing her first IVF attempt, without tubal problems, and without any prior pregnancies. The intercepts of the TM and our model corresponding to this new reference were −1.77 and −0.73, with associated LB rates of 15 and 33%. Our study provides supplementary evidence of the relevance of the TM in capturing the variability due to patient-mix, based on a few predictors presumably accounting for an important part of the maximum possible variability attributable to patient-mix, and remaining invariant and reproducible between centres. The poor performances of the TM observed in earlier validations are mainly due to the marked difference between the two intercepts of the model, essentially attributable to centre effect, and clearly not generalizable from one centre to another. Refining our model simply identified the adequate intercept, which immediately improved discrimination and calibration. In conclusion, given the considerable centre effect, external validation of a predictive model must not be limited to assessing the model alone, but must compare parameter estimates and adapt to the intercept.

Longitudinal linear trend in time

A time-dependent deterioration in the accuracy of the TM was recently hypothesized (Van Loendersloot et al., 2011), due to the influence of recent technological improvements in ART on LB rate. The observed decreasing C-statistics over time seem to support this hypothesis. A significant effect of a longitudinal linear trend was found, and no interaction was found between this trend and other predictors. Our results confirm the great need for a time trend correction. However, time does not seem to alter the influence of the other predictors, which appears to favour the hypothesis of a uniform increase in performance across the whole population and not for a subgroup of women in particular. When using the suggested reference population based on 30-year-old women, the difference in LB rates of 15 and 33% estimated from the TM and our data, respectively, may seem surprising. Bearing in mind that the data used for the original TM were collected before 1996, our model to predict LB in 1996 produced an estimate of LB = 0.17 (95% CI: 0.12–0.21), which includes the estimated value of 15% found with the TM. The intercept differences are essentially explained by a time trend in performance.

New predictors

The TM was constructed with the variables available in the HFEA data file. It might be suspected that other variables not documented in this file may have an additional effect. Among several variables studied in previous research, we investigated the effect of FSH (FSH < 10), normo-weighted women (18 < BMI < 26), and smoking habits, as surrogate predictors to the original TM. As suggested in a previous work (Van der Steeg et al., 2008), a BMI value over 29 was seen to reduce the chance of pregnancy by 4%, and severely obese women (BMI > 35) have 23–43% less chance of achieving pregnancy compared with women with BMI < 29. In our study by adjusting for the other predictors, normo-weighted women (18 < BMI < 26) were found to have higher LB (1.29; 95% CI: 1.14–1.4) compared with under- and over-weighted women, which corresponds to 25% more chance of becoming pregnant, assuming a mean LB rate around 20%.

Various authors underlined the predictive negative value of FSH at baseline (Sharif et al., 1998; Abdalla and Thum, 2004; Ferrari et al., 2011), particularly for older maternal age, in spite of its interaction with estradiol (Broekmans et al., 2006; Lee et al., 2009a,b; Grynberg and Fanchin, 2011). In our study, FSH values > 10 were associated with a negative prognosis (OR = 2.41), and women with FSH values below 10 were twice as likely to have an LB than those with FSH above 10.

Some studies suggested the negative effect of smoking habits (Augood et al., 1998). In our study, observed negative prognosis (OR = 0.60) for smokers, corresponded to the expected 29% reduction in LB predicted for past/current smokers compared with non-smokers.

These three new predictors were added to constitute our final model. Due to missing data for these variables, we assessed the consistency of results through a sensitivity analysis using missing data imputation techniques. We demonstrated that these supplementary variables provided a significant increase in discrimination (C = 0.71 instead of 0.69), but above all, we concluded that addition of FSH, smoking habits and BMI improved classification for a net 17.8% of individuals with events, without net loss for non-events.

This step aimed essentially at increasing the accuracy of the model; however, our findings on possible additional predictors in the TM can only be considered as hypothetical and require external validation in other centres.

Model performance and further improvements

After successive improvement stages, our final model can be considered as having the minimum acceptable quality for use in every day practice. Minimum acceptable discrimination was reached (C > 0.7, Swets, 1988) and calibration of the updated model was excellent. All the predictors of this model are easily available when starting a new cycle. A recent mono-centre attempt (Cai et al., 2011) at LB modelling achieved better discrimination (C = 0.77) where the quality of embryos was the most predictive variable. However, using this model is not applicable, as the quality of embryos is not often known when a woman starts a new cycle.

This model is now currently used in our centre through computer software based on a dialog box facilitating data entry of patient-mix. Our model is constantly evaluated by an adaptive statistical technique, accounting for new patients, and adding and testing new potential predictors.

Conclusion and implication for practice

Due to very marked inter-centre variability, a universal LB model will be generally characterized by poor predictive performances, although fitting the TM to one centre in particular provides a much better estimate. The ultimate practical purpose of a model is to provide the most accurate centre-specific prediction.
For centres interested in predicting LB, this study concludes that strict application of a model originating in other centres must be avoided because of expected poor performance. Our results show a simple way to develop a reasonably accurate centre-specific predictive model.

Despite poor predictive performances consistently confirmed by all external validations, we provided evidence that the TM is the undisputed model of reference, capturing a high proportion of the patient-mix effect. Model parameters remain invariant between centres, and as suggested by this study, across various cultures, countries and types of hospitals. The only difference we found was the intercept value, interpreted as the measurement of the global performance of one centre in particular for a population of reference.

The TM should be fitted to centre data. Fitting a logistic regression is an easy task, including complementary techniques to preserve stability and reduce the overfit of the model. The comparison of each estimate with the original TM is useful to assess whether the centre data confirm the general patient-mix trend. A difference is possible and should suggest that the study centre has developed expertise which is better adapted to particular subgroups compared with the TM standard.

For centres without statistical resources, we suggest an approximate method: assuming that the TM parameters of patient-mix remain rather constant between centres, and only the intercept varies, the overall LB rate will be calculated and used, under a slight transformation, as the intercept of the model. Although sub-optimal, this technique may be useful in providing a good approximation, particularly for centres only starting their data collection.

Our study provides some evidence that the use of a reasonably accurate predictive model at the centre level is now a relatively easy and affordable task. Based on its direct purpose as a predictive tool to decide whether to start an IVF cycle or not, such a model could be used extensively in statistical quality control, either when comparing LB performances in time within a centre or when comparing LB performances in different centres. Intra- and inter-centre comparisons involve heterogeneous populations with respect to patient-mix, so appropriate adjustment is required to provide a relevant comparison and increase statistical power. Beyond the scope of this paper, these considerations are worth further development, and are the objective of our future research.

Acknowledgements

The authors would like to thank Dr Etienne Varlan and Dr Carolina Colella (Merck Serono France) for their helpful discussion of the manuscript, and Dr Alain Platel (APMW) for editing.

Authors’ roles

P.A. was involved in study conception and study design, and writing the manuscript. P.L. was involved in data management and analysis, interpretation and writing the manuscript. A.G.L. contributed to the study conception and manuscript review.

Funding

No external funding was either sought or obtained for this study.

Conflict of interest

There were no commercial relationships (i.e. consultancies, patent-licensing agreements) that might pose a conflict of interest in connection with the submitted manuscript. The objective of this research was not directed toward any treatment effects.

References


Chuang CC, Chen CD, Chao KH, Chen SU, Ho HN, Yang YS. Age is a better predictor of pregnancy potential than basal follicle-stimulating hormone levels in women undergoing in vitro fertilization. Fertil Steril 2003; 79:63–68.


Stolwijk AM, Wetzels AM, Braat DD. Cumulative probability of achieving an ongoing pregnancy after in-vitro fertilization and intracytoplasmic sperm injection according to a woman’s age, subfertility diagnosis and primary or secondary subfertility. Hum Reprod 2000; 15:203–209.


