Predicting the cumulative chance of live birth over multiple complete cycles of in vitro fertilisation: an external validation study

Journal:	*Human Reproduction*
Manuscript ID	HUMREP-18-0281.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	28-Jun-2018
Complete List of Authors:	Leijdekkers, Jori; University Medical Centre Utrecht, Utrecht University, Reproductive Medicine and Gynaecology Eijkemans, Marinus; University Medical Centre Utrecht, Utrecht University, Julius Centre for Health Sciences and Primary Care Tilborg, Theodora; University Medical Centre Utrecht, Utrecht University, Reproductive Medicine and Gynaecology Oudshoorn, Simone; University Medical Centre Utrecht, Utrecht University, Reproductive Medicine and Gynaecology McLernon, David; University of Aberdeen, Division of Applied Health Sciences Bhattacharya, Siladitya; Cardiff University School of Medicine, School of Medicine, College of Biomedical and Life Sciences Mol, Ben; Monash University, Department of Obstetrics and Gynaecology Broekmans, Frank; University Medical Centre Utrecht, Utrecht University, Reproductive Medicine and Gynaecology Torrance, H; University Medical Centre Utrecht, Utrecht University, Reproductive Medicine and Gynaecology
Keywords:	Prediction model, External validation, Live birth, IVF/ICSI, INFERTILITY
Subject Section:	

TITLE PAGE

Title: Predicting the cumulative chance of live birth over multiple complete cycles of in vitro fertilisation: an external validation study

Running title: Validation of an IVF model predicting live birth

Authors:

J.A. Leijdekkers[1,*], M.J.C. Eijkemans[2], T.C. van Tilborg[1], S.C. Oudshoorn[1], D.J. McLernon[3], S. Bhattacharya[4], B.W.J. Mol[5], F.J.M. Broekmans[1], H.L. Torrance[1], on behalf of the OPTIMIST group.

[1]Department of Reproductive Medicine and Gynaecology, University Medical Centre Utrecht, Utrecht University, PO box 85500, 3508 GA Utrecht, The Netherlands. [2]Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, PO box 85500, 3508GA Utrecht, The Netherlands. [3]Institute of Applied Health Sciences, Medical Statistics Team, University of Aberdeen, Aberdeen AB25 2ZD, UK. [4]School of Medicine, College of Biomedical and Life Sciences, Cardiff University School of Medicine, Cardiff CF14 4XN, UK [5]Department of Obstetrics and Gynaecology, Monash University, VIC 3800 Clayton, Australia.

***Correspondence address:** J.A. Leijdekkers, Department of Reproductive Medicine and Gynaecology, University Medical Centre Utrecht, Utrecht University, PO box 85500, 3508 GA Utrecht, The Netherlands. E-mail: j.a.leijdekkers@umcutrecht.nl

ABSTRACT

Study question

Are the published pre-treatment and post-treatment McLernon models, predicting cumulative live birth rates (LBR) over multiple complete IVF cycles, valid in a different context?

Summary answer

With minor recalibration of the pre-treatment model, both McLernon models accurately predict cumulative LBR in a different geographical context and a more recent time period.

What is known already

Previous IVF prediction models have estimated the chance of a live birth after a single fresh embryo transfer, thereby excluding the important contribution of embryo cryopreservation and subsequent IVF cycles to cumulative LBR. In contrast, the recently developed McLernon models predict the cumulative chance of a live birth over multiple complete IVF cycles at two certain time points: a) before initiating treatment using baseline characteristics (pre-treatment model) and b) after the first IVF cycle adding treatment related information to update predictions (post-treatment model). Before implementation of these models in clinical practice, their predictive performance needs to be validated in an independent cohort.

Study design, size, duration

External validation study in an independent prospective cohort of 1515 Dutch women who participated in the OPTIMIST study (NTR2657) and underwent their first IVF treatment between 2011 and 2014. Participants underwent a total of 2881 complete treatment cycles, with a complete cycle defined as all fresh and frozen thawed embryo transfers resulting from one episode of ovarian stimulation. The follow up duration was 18 months after inclusion, and the primary outcome was ongoing pregnancy leading to live birth.

Participants/materials, setting, methods

Model performance was externally validated up to three complete treatment cycles, using the linear predictor as described by McLernon et al. to calculate the probability of live birth. Discrimination was expressed by the c-statistic and calibration was depicted graphically in a calibration plot. In contrast to the original model development cohort, anti-Müllerian hormone (AMH), antral follicle count (AFC) and body weight were available in the OPTIMIST cohort, and evaluated as potential additional predictors for model improvement.

Main results and the role of chance

Applying the McLernon models to the OPTIMIST cohort, the c-statistic of the *pre-treatment model* was 0.62 (95% confidence interval (CI) 0.59-0.64) and of the *post-treatment model* 0.71 (95% CI 0.69-0.74). The calibration plot of the *pre-treatment model* indicated slight overestimation of the cumulative LBR. To improve calibration, the *pre-treatment model* was recalibrated by subtracting 0.35 from the intercept. The *post-treatment model* calibration plot revealed accurate cumulative LBR predictions. After addition of AMH, AFC and body weight to the McLernon models, the c-statistic of the *updated pre-treatment model* improved slightly to 0.66 (95% CI 0.64-0.68), and of the *updated post-treatment model* remained at the previous level of 0.71 (95% CI 0.69-0.73).

Using the *recalibrated pre-treatment model*, a woman aged 30 years with two years of primary infertility who starts ICSI treatment for male factor infertility has a chance of 40% of a live birth from the first complete cycle, increasing to 72% over three complete cycles. If this woman weighs 70 kilograms, has an AMH of 1.5 ng/mL and an AFC of 10 measured at the beginning of her treatment, the *updated pre-treatment model* revises the estimated chance of a live birth to 30% in the first complete cycle and 59% over three complete cycles. If this woman then has 5 retrieved oocytes, no embryos cryopreserved and a single fresh cleavage stage embryo transfer in her first ICSI cycle, the *post-treatment model* estimates the chances of a live birth at 28% and 58%, respectively.

Limitations, reasons for caution

Two randomised controlled trials (RCT) evaluating the effectiveness of gonadotropin dose individualisation on basis of the AFC were nested within the OPTIMIST study. The strict dosing regimens, the RCT in- and exclusion criteria and the limited follow up time of 18 months might have influenced model performance in this independent cohort. Also, consistent with the original model development study, external validation was performed using the optimistic assumption that the cumulative LBR in couples who discontinue treatment without a live birth would have been equal to that of those who continue treatment.

Wider implications of the findings

After national recalibration to account for geographical differences in IVF/ICSI treatment, the McLernon prediction models can be introduced as new counselling tools in clinical practice to inform patients and to complement clinical reasoning. These models are the first to offer an objective and personalised estimate of the cumulative probability of live birth over multiple complete IVF cycles.

Study funding/competing interest(s):

No external funds were obtained for this study. M.J.C.E., D.J.M. and S.B. have nothing to disclose. J.A.L, S.C.O, T.C.v.T. and H.LT. received an unrestricted personal grant from Merck BV. B.W.M. is supported by a NHMRC Practitioner Fellowship (GNT1082548) and reports consultancy for ObsEva, Merck and Guerbet. F.J.M.B. receives monetary compensation as a member of the external advisory board for Merck BV (the Netherlands) and Ferring pharmaceutics BV (the Netherlands), for consultancy work for Gedeon Richter (Belgium) and Roche Diagnostics on automated AMH assay development, and for a research cooperation with Ansh Labs (USA).

Trial registration number

Not applicable

KEYWORDS

Prediction model, external validation, live birth, IVF/ICSI, infertility, cumulative live birth, personalised, counselling, prognostic research

Introduction

Infertility is defined as the failure to conceive within 12 months of regular unprotected intercourse, and affects approximately one in six couples (Oakley *et al.*, 2008; Zegers-Hochschild *et al.*, 2017). The majority of infertile couples seek fertility care, and many of those with prolonged unresolved infertility will be treated with ART regardless of cause (Boivin *et al.*, 2007; Datta *et al.*, 2016). IVF and ICSI are both widely used techniques for couples with infertility. Globally more than 1.6 million annual cycles of IVF/ICSI are performed and while success rates have increased over time (Dyer *et al.*, 2016; McLernon *et al.*, 2016), this treatment is still not effective for all infertile couples, with live birth rates (LBR) at around 25-30% per treatment cycle (Malizia *et al.*, 2009; McLernon *et al.*, 2016; de Neubourg *et al.*, 2016). Since IVF/ICSI is expensive and carries several risks, the probability of a live born child should be weighed against the risks and costs of this treatment.

Several prognostic models have been developed to objectively estimate the probability of a live birth after IVF/ICSI treatment (Leushuis *et al.*, 2009; van Loendersloot *et al.*, 2014). It is known that prediction models often perform optimistically in their development sample, even after correction by internal validation. This is caused by overfitting, which occurs when the model corresponds too closely to the development data due to the inclusion of too many predictors (Moons, Kengne, Woodward, *et al.*, 2012). External validation in an independent cohort of women is thus essential to examine the performance and generalisability of the prediction model (Altman *et al.*, 2009; Harrell *et al.*, 1996). Unfortunately, most of the currently available models that predict the chance of a live birth after IVF/ICSI treatment have never been externally validated (Leushuis *et al.*, 2009; van Loendersloot *et al.*, 2014). Also, the majority of these models predict the probability of a live birth after a single fresh embryo transfer, excluding the important contribution of embryo cryopreservation and subsequent treatment cycles to LBR. This limits their potential as counselling tools for couples and clinicians, especially considering the increased use and improved techniques of embryo cryopreservation and frozen thawed embryo transfer cycles in recent years (Wong *et al.*, 2014).

Three of the largest model development studies for prediction of live birth after IVF and/or ICSI treatment used data from the Human Fertilisation and Embryology Authority (HFEA) database in the UK (McLernon *et al.*, 2016; Nelson and Lawlor, 2011; Templeton *et al.*, 1996). Treatment and outcome data from all licenced fertility clinics within the UK have been recorded in this database since 1992. The two models developed by Templeton et al. and Nelson et al. were both externally validated, and their predictive performance was compared to one another in several studies (Arvis *et al.*, 2012; van Loendersloot *et al.*, 2011; Smeenk *et al.*, 2000; Smith *et al.*, 2015; te Velde *et al.*, 2014). Although these models have been recommended in previous studies and used internationally to predict live birth after IVF and ICSI (Leushuis *et al.*, 2009; Smith *et al.*, 2015; te Velde *et al.*, 2014), neither model predicts cumulative LBR over multiple IVF/ICSI treatment cycles including frozen thawed embryo transfer cycles.

Recently, a new model was developed by McLernon et al. using the HFEA database (McLernon *et al.*, 2016). This model is the first to provide an individualised estimate of the cumulative chance of a live birth over multiple complete cycles of IVF/ICSI, with a complete cycle defined as all fresh and frozen thawed embryo transfers resulting from one episode of ovarian stimulation. For model development, data from 113 873 women and 184 269 complete cycles between 1999 and 2009 were used. Internal validation of the model showed promising results, however evaluation of the predictive performance of the model in a different geographical context using more contemporary data has yet to be performed. Additionally, a number of potential key predictors, such as measures for ovarian reserve and female body weight, were unavailable in the HFEA database and could not be included in the original model (McLernon *et al.*, 2016).

The main objective of the current study was therefore to perform geographical and temporal validation of the new HFEA model by using recent data from a different country. We also wanted to determine whether inclusion of additional parameters, such as female body weight and ovarian reserve test results i.e. antral follicle count (AFC) and anti-Müllerian hormone (AMH), could improve the predictive performance of the model.

Materials and methods

Data sources

External validation was performed on data from the OPTIMIST study (van Tilborg, Oudshoorn, *et al.*, 2017). This multicentre prospective cohort study included 1515 women from 25 infertility centres in the Netherlands between May 2011 and May 2014. Participants were younger than 44 years of age, had regular menstrual cycles and no significant uterine or ovarian abnormalities on transvaginal ultrasound. Women with polycystic ovarian syndrome, metabolic or endocrine abnormalities or undergoing oocyte donation were excluded. All participants were included before their first IVF/ICSI cycle, or the first cycle after a previous live birth. The primary outcome was ongoing pregnancy, achieved within 18 months of follow up, and resulting in live birth. Ethical approval for the OPTIMIST study was obtained from the Institutional Review Board of the University Medical Centre Utrecht (MEC 10-273), and all participants provided written informed consent. A more detailed description of study procedures and results were reported previously (Oudshoorn *et al.*, 2017; van Tilborg *et al.*, 2012; van Tilborg, Oudshoorn, *et al.*, 2017; van Tilborg, Torrance, *et al.*, 2017).

McLernon model

The McLernon model consists of two clinical prediction models to estimate the individualised cumulative chance of a live birth over a maximum of six complete treatment cycles. **Before initiating treatment**, the *pre-treatment model* predicts the probability of a live birth from both fresh and frozen thawed embryo transfers based on couple characteristics and the use of IVF or ICSI. Included predictors are: female age (years), duration of infertility (years), previous pregnancy, causes of infertility (tubal factor, anovulation, male factor, unexplained infertility), type of treatment (IVF or ICSI) and treatment year (see Supplementary Text 1).

After the first fresh treatment cycle, treatment specific characteristics from this cycle are added in the *post-treatment model* to update the predicted probability. Added predictors are: number of oocytes, cryopreservation of embryos, and the number and stage of embryos at the first fresh embryo transfer (single, double or triple embryo transfer; blastocyst or cleavage stage). All causes of infertility are excluded as predictors in the post-treatment model, except for tubal factor (see Supplementary Text 2). For women with zero oocytes collected in the first cycle, a separate post-treatment model is available.

To predict the probability of a live birth in the *i*th cycle, assuming no live birth occurred in the previous cycle(s), complete cycle number is included in both models as a discrete time variable. A complete cycle includes all fresh and frozen thawed embryo transfers resulting from one episode of ovarian stimulation. With the predicted probability of a live birth per subsequent complete cycle, the cumulative probability of a live birth can be calculated up to six complete cycles (see Supplementary Text 1 and 2).

Statistical analysis

Nine predictor variables had missing values (Table I). The proportion of missing values was low (< 2.5%), except for AMH (11.2%). During the OPTIMIST study, blood sampling was performed on the day of randomisation. Logistic issues prevented blood sampling in some cases, thus compromising the ability to undertake post-hoc measurements of AMH in the total population. As the reasons for missing values were considered to be unrelated to the AMH value itself or the measurement, these were defined as missing (completely) at random.

Multiple imputation was applied for predictors with missing values in the OPTIMIST database (Sterne *et al.*, 2009). In this process 10 imputed datasets were created using a multivariate imputation by chained equations (MICE) algorithm (van Buuren and Groothuis-Oudshoorn, 2011). Predicted probabilities for a live birth were calculated on each imputed dataset, using the predictors and parameter-estimates of both the pre-treatment model as well as the post-treatment model as described by McLernon et al 2016 (McLernon *et al.*, 2016). In accordance with the original models, the variables female age, treatment year and number of oocytes were treated with restricted cubic splines in the validation process. The separate post-treatment model for women with zero oocytes collected in the first treatment cycle was not validated in this study, as the number of women for this analysis was too low in the OPTIMIST database. Cumulative probabilities were calculated up to three complete IVF/ICSI cycles, as most couples in the Netherlands only have three treatment cycles due to the current reimbursement policy. Also, the OPTIMIST follow up period was 18 months, reducing the number of women with more than three treatment cycles. The validation process was performed ten times on each of the imputed datasets and separate results were pooled using Rubin's rules (Rubin, 2004).

The predictive performance of the McLernon models was evaluated in terms of discrimination and calibration. Discrimination quantifies the ability of a model to correctly differentiate between subjects with an event and subjects without an event (Moons, Kengne, Woodward, *et al.*, 2012). In the context of fertility treatment, it is the ability of the models to distinguish between women with a live birth and women without a live birth after IVF/ICSI treatment. It is expressed by the c-statistic or the area under the receiver operating curve (AUROC), which ranges between 0.5 and 1. A c-statistic of 1 indicates perfect discrimination, whereas a c-statistic of 0.5 represents a model with no discrimination at all. In this study, the c-statistic (and 95% CI) was calculated using the method suggested by Harrell et al. (Harrell *et al.*, 1996).

Calibration describes the degree of agreement between predicted probabilities and observed outcomes (Moons, Kengne, Woodward, *et al.*, 2012), in this context the predicted probability of a live birth and the observed LBR. Calibration can be assessed graphically by forming subgroups of patients determined by ranges of predicted probabilities, and then plotting the observed proportion of events against the mean predicted probability within these subgroups. When perfect calibration is present, the plot shows a diagonal line with a slope of one and an intercept of zero. In the current study, five equal subgroups of patients were formed. This was based on the sample size of the OPTIMIST cohort and the related precision of the point estimates in the calibration plot. Within these subgroups, the Kaplan Meier estimates of the observed cumulative LBR over three complete treatment cycles were plotted against the mean predicted probability of cumulative live birth. A smoothed line was then added in this plot using the proportional hazard regression approach described by Harrell et al (Harrell *et al.*, 1996). In addition to this, a systematic difference in the predicted and observed LBR was assessed by using calibration-in-the-large (Steyerberg, 2009), and the intercept of the prediction models was adjusted in case a systematic over- or underestimation was present.

Updating the models

Following the external validation of the models, the additional value of updating the McLernon models with pre-specified new biomarkers was evaluated. AMH (ng/mL), AFC (2-10 mm) and body weight (kg) were added to the pre-treatment and post-treatment model in a multivariable logistic regression analysis, in which the linear predictor of the McLernon model was entered as a fixed variable. The final model was established using a manual backward selection process. Predictors were eliminated from the model according to the Akaike Information Criterion (AIC) (Akaike, 1974).

The predictive performance of the new updated models was evaluated by calculating the c-statistic (and 95% CI). To assess for overfitting, internal validation was performed by bootstrapping (Steyerberg, 2009). Two hundred bootstrap samples, all of which were of the same size as the original validation sample, were created by random sampling with replacement (Harrell, 2001; Steyerberg, 2009). In each bootstrap sample, a new model was fitted with the same predictors as the updated models. The c-statistic was calculated for each of the 200 sample derived models, in both the bootstrap sample as well as the original validation cohort. The difference between these two c-statistics was calculated for each of the 200 sample derived models, and averaged to give the optimism estimate. This was subtracted from the original c-statistic to obtain the optimism corrected c-statistic for the updated models.

All statistical analyses were performed using R for Windows (version 3.3.2; R Foundation for Statistical Computing, Vienna, Austria).

Results

Of the 1515 women included in the OPTIMIST study, four were excluded in the current study as they never started IVF/ICSI treatment. A total of 2881 IVF/ICSI cycles were performed over a period of 18 months of follow up. Table I shows the patient and first cycle treatment characteristics of the OPTIMIST cohort (validation sample) and the HFEA cohort (development sample). Women included in the validation sample were about the same age as women in the development sample, but had a shorter average duration of infertility. The causes of infertility showed a similar distribution across both samples, with the exception of anovulation which rendered women ineligible for the OPTIMIST study. The treatment characteristics showed that embryo cryopreservation was more frequently performed after the first IVF/ICSI cycle in the validation sample and that these women most often had a cleavage stage single embryo transfer in the first fresh cycle, whereas women in the development sample most often had a cleavage stage double embryo transfer. No formal assessment was performed for the differences and similarities between the cohorts, as a description rather than a p-value is considered to be useful for interpretation of the models' performance in this external validation study.

The flowchart in Figure 1 shows the number of women in the OPTIMIST and HFEA cohorts who started a treatment cycle, had a live birth or discontinued treatment without having a live birth. The LBR per cycle was similar in both cohorts for the first, second and fourth treatment cycle. In the third cycle the LBR was slightly higher in the OPTIMIST cohort compared to the HFEA cohort. As few women in the OPTIMIST cohort received a fifth or sixth cycle, LBR in these cycles could not be compared. The proportion of women without a live birth that continued treatment was higher after the first and second cycle in the OPTIMIST cohort as compared to the HFEA cohort. After the third cycle, the proportion continuing treatment in the OPTIMIST cohort decreased, while it remained constant in the HFEA cohort. At the end of follow up, 52% of the women in the OPTIMIST study had a treatment related live birth. The overall LBR of the HFEA cohort was 43% over six complete IVF/ICSI cycles.

As mentioned previously, external validation of the McLernon models was performed up to three complete treatment cycles, and therefore the fourth, fifth and sixth complete treatment cycle in the OPTIMIST dataset (n=102 complete treatment cycles, n= 15 live births) were excluded from further analysis. Also, for the post-treatment model validation, women with zero oocytes collected in the first treatment cycle were excluded (n= 226 women, n = 526 complete treatment cycles, n= 82 live births) as a separate model was developed for this group of women by McLernon et al (McLernon *et al.*, 2016). Due to the small numbers, this separate model could not be validated in this study.

Discrimination and calibration

In the validation sample, the pooled c-statistic for the pre-treatment model was 0.62 (95% CI 0.59-0.64) and for the post-treatment model 0.71 (95% CI 0.69-0.74). Figure 2a and 3 show the calibration plots for both original models, depicting the correlation between the observed and predicted cumulative LBR. The pre-treatment calibration plot had an intercept of -0.23 (95% CI -0.36- -0.10) and a slope of 0.98 (95% CI 0.69-1.27), and the post-treatment calibration plot had an intercept of -0.01 (95% CI -0.12-0.11) and a slope of 0.97 (95% CI 0.77-1.19).

The pre-treatment model systematically overestimated the cumulative LBR over three complete cycles for women in the validation sample. This is shown by a calibration curve with most of the confidence intervals under the reference line (Figure 2a), indicating significantly higher predicted probabilities than observed LBR. The calibration-in-the-large analysis confirmed this systematic overestimation with an intercept of -0.35. To improve calibration, the pre-treatment model was thus adjusted by subtracting 0.35 from the intercept of the original linear predictor, which decreased the predicted odds of a live birth by a factor of 1.42 (see Supplementary Text 3). The calibration plot of the recalibrated pre-treatment model showed improved accuracy of the predictions, with all confidence intervals overlapping the reference line (Figure 2b). In contrast to the pre-treatment model, the post-treatment model correctly estimated the cumulative LBR in the validation sample, as is shown by a calibration plot with confidence intervals overlapping the reference line indicating no significant over- or underestimation (Figure 3).

Updating of the models

Addition of the biomarkers AMH, AFC and body weight to the pre-treatment and post-treatment model in a multivariable regression analysis resulted in two new updated models. The updated pre-treatment model included all three biomarkers as additional predictors for live birth. Since the relationship between both AMH and AFC with the probability of live birth was non-linear, these predictors were included using restricted cubic splines (see Supplementary Figure 1). The updated post-treatment model included only AFC and AMH as additional predictors for live birth, of which AFC was modelled by using restricted cubic splines (see Supplementary Figure 2). After internal validation of the updated models by bootstrapping, the updated pre-treatment model had a corrected c-statistic of 0.66 (95% CI 0.64-0.68) and the updated post-treatment model had a corrected c-statistic of 0.71 (95% CI 0.69-0.73). The addition of AFC, AMH and body weight thus resulted in a slight improvement of the discriminatory capacity of the pre-treatment model, while addition of AFC and AMH had no beneficial effect on the discriminative performance of the post-treatment model.

Examples of model predictions

Figures 4, 5 and 6 show examples of model predictions as illustration for clinical application. Figure 4 presents predictions of the *recalibrated pre-treatment model* for couples with primary infertility caused by a male factor. Cumulative probabilities of live birth are calculated up to three complete ICSI cycles, and are differentiated by female age (30 or 40 years) and duration of infertility (2 years or 5 years). As is shown in figure 4, age is the most important predictor in the pre-treatment model. A 30-year-old woman with 2 years of infertility has a predicted probability of a live birth of 0.40 in the first ICSI cycle, increasing to 0.72 over three complete cycles. For a 40-year-old woman with 2 years of infertility, these probabilities are 0.15 and 0.32 respectively.

Figure 5 shows predictions of the *updated pre-treatment model*, with AMH, AFC and body weight as new predictors in the model. Predictions are presented for couples with two years of primary infertility caused by a male factor, and differentiation is based on female age (30 or 40 years), AMH (2.0 or 0.5 ng/mL) and AFC (15 or 7). In all scenarios the female body weight is 70 kilograms. A 30-year-old woman with an average ovarian reserve at the start of her first treatment – indicated by an AMH of 2.0 ng/mL and an AFC of 15 – has a predicted probability of a live birth of 0.37 in the first cycle and 0.69 over three cycles (0.17 and 0.37 for a 40-year-old woman). If this woman has a reduced ovarian reserve – indicated by an AMH of 0.5 ng/mL and an AFC of 7 – the predicted probabilities decrease to 0.19 and 0.42, respectively (0.08 and 0.18 for a 40-year-old woman).

Figure 6 shows predictions of the *post-treatment model,* which revises the predicted probabilities of the pre-treatment models by adding information of the first treatment cycle. Predictions are calculated for women with two years of primary, non-tubal infertility and are differentiated by female age (30 or 40 years), number of oocytes (10 or 5) and embryo cryopreservation (yes or no). In all scenarios the woman received a cleavage stage single embryo transfer. The predicted probabilities of a live birth for women with a favourable prognosis – aged 30-years, 10 oocytes retrieved and cryopreserved embryos – is 0.49 in the first ICSI cycle, increasing to 0.83 over 3 complete cycles. In contrast, for women with a poorer prognosis – aged 40 years, 5 oocytes retrieved and no embryos cryopreserved – the predicted probabilities are 0.11 and 0.26, respectively.

Discussion

Main findings

This external validation study of the McLernon pre-treatment and post-treatment model found that, after minor recalibration of the intercept of the pre-treatment model, both models accurately predict the cumulative probability of live birth up to three complete IVF/ICSI cycles in a more contemporary cohort in another country. The discriminatory capacity of the pre-treatment model in an external cohort was limited, whereas the post-treatment model had a fair ability to discriminate between couples with and without a live birth after treatment.

Strengths

This study focuses on the external validation of an IVF prediction model, which is an essential but frequently overlooked step before implementation of a prediction model in clinical practice (Altman *et al.*, 2009). In contrast to redeveloping new models for the same outcome, external validation and updating of existing models prevents the loss of scientific information by combining the information captured in the original model with information of a new patient cohort (Moons, Kengne, Grobbee, *et al.*, 2012).

Embryo cryopreservation has become an important part of IVF/ICSI treatment, and most couples have more than just one complete treatment cycle (Wong *et al.*, 2014). Unlike previous prediction models (Leushuis *et al.*, 2009; van Loendersloot *et al.*, 2014), the McLernon models provide a more useful estimate of cumulative treatment success. As such, the validation of these models represents a significant step forward in creating a clinically useful tool to manage expectations and to inform decision making around IVF.

This study benefits from the prospective design of the OPTIMIST study, which has ensured reliable data collection, with relatively low numbers of missing values and a low risk of selection bias. The multicentre design resulted in a highly representable cohort for Dutch fertility care. And although it is known that the

IVF/ICSI success rates vary between fertility centres, the inclusion of multiple centres will increase the generalisability and applicability of the external validation of the McLernon models within the Netherlands.

Furthermore, the external validation was performed on data collected in a recent time period (2011-2014). Due to changing patient populations, new treatment protocols, improving technologies and increasing success rates over time, prediction models in reproduction medicine have no static form and should be regularly updated to optimally reflect the latest circumstances in which they are used (Altman *et al.*, 2009). As the McLernon models were developed on data collected between 1999 and 2009, data of the more recently performed OPTIMIST study were helpful to investigate if model performance was still accurate in current practice.

Weaknesses

This study has a number of limitations. First, the external validation involved data from a prospective cohort study within which two randomised controlled trials were embedded evaluating the effectiveness of individualised doses of gonadotropins based on AFC. Strict dosing regimens might have affected some treatment outcomes, such as cancellation rates and number of oocytes, thus influencing the predictive capacity of the models in the validation sample. However, as the OPTIMIST study found no difference between the dosing regimens on cumulative live birth rates, the impact on model performance is likely to be minimal.

Second, the OPTIMIST study used strict eligibility criteria. Therefore, the validation sample does not fully represent the diversity of the patient population initiating IVF/ICSI treatment in the Netherlands. As none of the women in the validation sample were anovulatory, external validation of the models was only performed for an ovulatory population. This limits the generalisability of the models to some extent, as the original McLernon models were developed in a population which also included anovulatory women. Also, it could have had some impact on model performance. However, since anovulation had only a small predictive value in the pre-treatment model, and the majority of couples underwent IVF/ICSI for other indications, a large impact on model performance is unlikely.

Third, the OPTIMIST study had a follow up period of 18 months, leading to small numbers of women with more than three complete treatment cycles. Model performance could therefore only be reliably validated up to three complete cycles. However, most couples in the Netherlands complete a maximum of three treatment cycles which is partly due to the national reimbursement policy, but also by the high rates of embryo cryopreservation, increasing the number of embryo transfers and LBR per cycle. Therefore, model validation up to three complete cycles has particular clinical relevance for current Dutch fertility care.

Last, the original McLernon prediction models were developed on linked cycle data, which were then used to estimate cumulative pregnancy chances. Therefore, these models used the optimistic assumption that the cumulative LBR in couples who discontinue IVF treatment without a live birth would have been equal to that of couples who continue further treatment cycles, after correction of predictor effects. This assumption tends to lead to overestimation of the cumulative LBR, as women with a low prognosis of achieving a live birth are generally more likely to discontinue treatment (Brandes *et al.*, 2009; Olivius *et al.*, 2004). Since the reasons for treatment withdrawal were unknown in the current external validation study, a similar method was used that probably resulted in some degree of overestimation of the cumulative LBR in the validation cohort. However, as the original McLernon models were developed with this approach, and the predictions for cumulative LBR over multiple complete cycles were considered to be clinically more relevant than per cycle predictions, we feel that the current method is the best option for the external validation of the McLernon models.

Explanation of findings

The discriminatory capacity of the pre-treatment model was markedly lower in the validation sample than in the development sample. In the development study, a c-statistic of 0.73 (95% CI 0.72-0.74) was reported, whereas the present study found a c-statistic of 0.62 (95% CI 0.59-0.64). For the post-treatment model, the discriminatory performance in the validation sample was comparable to that in the development sample, with a c-statistic of 0.71 (95% CI 0.69-0.74) and 0.72 (95% CI 0.71-0.73) respectively (McLernon *et al.*, 2016). As it is known that prediction models tend to perform too optimistically in the development dataset due to overfitting, some reduction in model performance is to be expected during external validation due to the differences between samples (Altman *et al.*, 2009; Moons, Kengne, Woodward, *et al.*, 2012). This, to some extent, also explains the lower overall performance of the pre-treatment model. The comparable performance of the post-treatment model in both samples indicates that the treatment related variables that were added to this model (number of oocytes, cryopreservation of embryos, and the number and stage of embryos) are important predictors for live birth after treatment.

Other than the influence of overfitting, some key differences between the Dutch and UK healthcare systems may also have affected the models' performance in this external validation study. An important factor is the reimbursement policy for fertility treatment. All Dutch infertile couples are insured for a minimum of three complete IVF/ICSI cycles. In contrast, most couples in the UK receive no standard funding for ART (Berg Brigham *et al.*, 2013). Since IVF/ICSI treatment is expensive, this induces discrepancies in the patient population initiating and continuing treatment between the two study samples (Rajkhowa *et al.*, 2006). As can be seen in the baseline table (Table I) and flowchart (Figure 1), couples in the UK had a longer average duration of infertility before starting treatment and were more likely to discontinue treatment after the first and second cycles than couples in the Netherlands. Also, the decrease in LBR is more evident in the UK than in the Netherlands over the first three cycles, which suggests that differences exist in both reasons for discontinuation as well as prognostic profiles of women discontinuing treatment in the two countries. These phenomena are, in part, financially driven, and could partially explain the difference in predictive ability of the UK models in the Dutch cohort.

Furthermore, despite the fact that the infertility guidelines of both countries include similar approaches for treatment of infertile couples, there are important variations in treatment characteristics between the two study samples (Dutch Society of Obstetrics and Gynaecology (NVOG), 2010; National Institute for Health and Care Excellence (NICE), 2013). Some of these differences are mainly due to changes in clinical practice over time. As is shown by the baseline table (Table 1), women in the more recent Dutch cohort (2011-2014) generally had a single embryo transfer in their first fresh treatment cycle, whereas women in the earlier UK cohort (1999-2009) most often had a double embryo transfer. Also, embryo cryopreservation was performed in over half of the Dutch women as compared to only a quarter of the women in the UK. Other differences are explained by variation in treatment protocols between geographic locations. For one, no blastocyst stage embryos transfers were performed in the Netherlands in contrast to the proportion of blastocyst stage embryo transfers in the UK of more than 10%. Also, Dutch women more frequently had no embryo available for transfer after their first treatment cycle, which is most likely caused by strict cancellation criteria particularly for hyper response. These differences in treatment characteristics suggest that the development sample does not fully reflect clinical practice in a more recent time period and in a different geographic context. As cumulative LBR are substantially affected by the variation in treatment characteristics (Glujovsky *et al.*, 2016; Pandian *et al.*, 2013; Wong *et al.*, 2014), this could explain part of the different performance of the pre-treatment model in the validation sample . The stable performance of the post-treatment model, which includes embryo stage and embryo cryopreservation as important predictors, seems to confirm the impact of the variation in these variables on model performance.

The addition of measures of ovarian reserve, i.e. AMH and AFC, and body weight to the McLernon prediction models revealed only a marginal improvement of model performance in the OPTIMIST dataset. The additional value of these tests can therefore be questioned, especially in view of the extra costs and physical burden on the patient. Female age is one of the most important predictors in the McLernon models (McLernon *et al.*, 2016). As female age is correlated with the ovarian reserve, adding

AMH and AFC provides limited new information to the prediction models. This is in line with previous studies that showed that ovarian reserve tests have no added value to the use of female age alone in the prediction of ongoing pregnancy after treatment (Broer *et al.*, 2013). Other potential predictors for live birth, such as ethnicity, smoking status and alcohol intake, were not included in this update of the McLernon model (Dhillon *et al.*, 2015; Rossi *et al.*, 2011; Waylen *et al.*, 2009). The additional value of these variables for model performance was considered uncertain, as the reporting is remarkably subjective and/or often incomplete (Liber and Warner, 2018; Stockwell *et al.*, 2016).

Clinical implications

Discrimination and calibration have been recognized as measures to evaluate the performance of prediction models (Altman *et al.*, 2009; Steyerberg, 2009). However, the discriminative ability at the binary level of most prediction models in reproductive medicine, as expressed by the c-statistic, is considerably low (Leushuis *et al.*, 2009). As at the moment of prediction the outcome of pregnancy has not yet occurred, the c-statistic is determined using the calculated probability of pregnancy. The maximum value of the c-statistic depends on the variability of these calculated probabilities in the infertile population. Since infertility is a complex and multifactorial health problem and due to the absence of strong predictors for live birth – particularly pre-treatment – , the probability distribution in infertile couples that have a live birth has a considerable overlap with the distribution of those without a live birth. Therefore the maximum c-statistic can be expected to be low (Cook, 2007; Coppus *et al.*, 2009), as is seen in the external validation of the pre-treatment model. However, this does not necessarily imply that such prediction models have limited use in clinical practice. Models with reliable predictions and a clinically useful distribution of probabilities for achieving a live birth, as assessed by calibration, can still support patients and clinicians in clinical decision making around infertility treatment (Coppus *et al.*, 2009).

As the calibration plots of both the recalibrated pre-treatment model and the post-treatment model indicate accurate predictions with a useful range of prognoses, these models can be used within the Netherlands as counselling tools to complement clinical reasoning at two certain time points. Before initiating treatment, the recalibrated pre-treatment model offers couples and clinicians a personalised and objective estimate of success over multiple complete treatment cycles. And after the first fresh embryo transfer, the post-treatment model provides a revised estimate using treatment related information to personalize the predictions even more. Despite the applicability of the models as counselling tools to inform patients about their prognosis, the McLernon models should not yet be used for decisions on whether or not to withhold fertility treatment. The impact of such model-based decisions on cost-benefit outcomes should be investigated first and proven to be beneficial. To implement the McLernon models as counselling tools in other countries as well, national recalibration is recommended to account for geographical differences in IVF/ICSI treatment.

The McLernon models were converted into an online calculator to facilitate the use of the models in clinical practice (https://w3.abdn.ac.uk/clsm/opis). As the original pre-treatment model overestimates cumulative LBR for couples in the Netherlands, conversion of the recalibrated pre-treatment model into a new online calculator is needed for implementation in Dutch clinical practice. This tailored online calculator can then provide accurate and up to date predictions for couples and clinicians in the Netherlands. Ultimately, the online calculator will be offered for implementation on the websites of the Dutch Patient Association for people with fertility problems ‘Freya’ and the Dutch Association of Obstetrics and Gynaecology (NVOG) to increase the accessibility of the models.

Research implications

Following this external validation study, future studies could focus on the impact of introducing the McLernon prediction models in clinical practice, and assess changes in patient and clinicians’ behaviour and its effects on LBR and cost-effectiveness.

In conclusion, after minor recalibration of the pre-treatment model, the McLernon models have proven to be valid in predicting the chance of cumulative live birth after multiple complete treatment cycles in another geographical context and in a more recent time period. Updating the models with AMH, AFC and body weight revealed only a marginal improvement of predictive performance. Following national recalibration, implementation of the McLernon models as counselling tools in clinical practice will provide infertile couples and clinicians with objective and personalized estimates of success over multiple complete IVF/ICSI cycles.

Acknowledgements

We would like to thank the women who participated in the OPTIMIST study and the staff of the participating hospitals for their contributions to the OPTIMIST study.

Authors' roles

T.C.v.T. and S.C.O and all other members from the OPTIMIST study group collected the data. D.J.M., S.B., F.J.M.B. and H.L.T were involved in study conception and study design. J.A.L. and M. J. C. E. performed the statistical analysis. J.A.L. drafted the manuscript. J.A.L., M.J.C.E., F.J.M.B. B.W.M., H.L.T interpreted the data. All authors participated to the discussion of the findings and revised the manuscript.

Funding

No external funding was obtained for this study.

Conflict of interest

M.J.C.E., D.J.M. and S.B. have nothing to disclose. J.A.L, S.C.O, T.C.v.T. and H.LT. received an unrestricted personal grant from Merck BV. B.W.M. is supported by a NHMRC Practitioner Fellowship (GNT1082548) and reports consultancy for ObsEva, Merck and Guerbet. F.J.M.B. receives monetary compensation as a member of the external advisory board for Merck Serono (the Netherlands) and Ferring pharmaceutics BV (the Netherlands), for consultancy work for Gedeon Richter (Belgium) and Roche Diagnostics on automated AMH assay development, and for a research cooperation with Ansh Labs (USA).

References

Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974;**19**:716–723.

Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;**338**:b605.

Arvis P, Lehert P, Guivarc'h-Leveque A. Simple adaptations to the Templeton model for IVF outcome prediction make it current and clinically useful. *Hum Reprod* 2012;**27**:2971–2978.

Berg Brigham K, Cadier B, Chevreul K. The diversity of regulation and public financing of IVF in Europe and its impact on utilization. *Hum Reprod* 2013;**28**:666–75.

Boivin J, Bunting L, Collins JA, Nygren KG. International estimates of infertility prevalence and treatment-seeking: potential need and demand for infertility medical care. *Hum Reprod* 2007;**22**:1506–12.

Brandes M, van der Steen JOM, Bokdam SB, Hamilton CJCM, de Bruin JP, Nelen WLDM, Kremer JAM. When and why do subfertile couples discontinue their fertility care? A longitudinal cohort study in a secondary care subfertility population. *Hum Reprod* 2009;**24**:3127–35.

Broer SL, van Disseldorp J, Broeze KA, Dolleman M, Opmeer BC, Bossuyt P, Eijkemans MJC, Mol B-WJ, Broekmans FJM, Broer SL, *et al.* Added value of ovarian reserve testing on patient characteristics in the prediction of ovarian response and ongoing pregnancy: an individual patient data approach. *Hum Reprod Update* 2013;**19**:26–36.

Cook NR. Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve. *Clin Chem* 2007;**54**:17–23.

Coppus SFPJ, van der Veen F, Opmeer BC, Mol BWJ, Bossuyt PMM. Evaluating prediction models in reproductive medicine. *Hum Reprod* 2009;**24**:1774–1778.

Datta J, Palmer MJ, Tanton C, Gibson LJ, Jones KG, Macdowall W, Glasier A, Sonnenberg P, Field N, Mercer CH, *et al.* Prevalence of infertility and help seeking among 15 000 women and men. *Hum Reprod* 2016;**31**:2108–2118.

de Neubourg D, Bogaerts K, Blockeel C, Coetsier T, Delvigne A, Devreker F, Dubois M, Gillain N, Gordts S, Wyns C. How do cumulative live birth rates and cumulative multiple live birth rates over complete courses of assisted reproductive technology treatment per woman compare among registries? *Hum Reprod* 2016;**31**:93–99.

Dhillon RK, Smith PP, Malhas R, Harb HM, Gallos ID, Dowell K, Fishel S, Deeks JJ, Coomarasamy A. Investigating the effect of ethnicity on IVF outcome. *Reprod Biomed Online* 2015;**31**:356–363.

Dutch Society of Obstetrics and Gynaecology (NVOG). Landelijke Netwerkrichtlijn Subfertiliteit. 2010. Available at: http://nvog-documenten.nl.

Dyer S, Chambers GM, de Mouzon J, Nygren KG, Zegers-Hochschild F, Mansour R, Ishihara O, Banker M, Adamson GD. International Committee for Monitoring Assisted Reproductive Technologies world report: Assisted Reproductive Technology 2008, 2009 and 2010. *Hum Reprod* 2016;**31**:1588–1609.

Glujovsky D, Farquhar C, Quinteiro Retamar AM, Alvarez Sedo CR, Blake D. Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology. *Cochrane Database Syst Rev* 2016:CD002118.

Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer-Verlag , 2001.

Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361–387.

Leushuis E, van der Steeg JW, Steures P, Bossuyt PMM, Eijkemans MJC, van der Veen F, Mol BWJ, Hompes PGA. Prediction models in reproductive medicine: a critical appraisal†. *Hum Reprod Update* 2009;**15**:537–552.

Liber AC, Warner KE. Has Underreporting of Cigarette Consumption Changed Over Time? Estimates Derived From US National Health Surveillance Systems Between 1965 and 2015. *Am J Epidemiol* 2018;**187**:113–119.

Malizia BA, Hacker MR, Penzias AS. Cumulative Live-Birth Rates after In Vitro Fertilization. *N Engl J Med* 2009;**360**:236–243.

McLernon DJ, Steyerberg EW, te Velde ER, Lee AJ, Bhattacharya S. Predicting the chances of a live birth after one or more complete cycles of in vitro fertilisation: population based study of linked cycle data from 113 873 women. *BMJ* 2016;**355**:i5735.

Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;**98**:691–8.

Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012;**98**:683–90.

National Institute for Health and Care Excellence (NICE). Fertility problems: assessment and treatment. Clinical guideline. 2013. Available at: https://www.nice.org.uk.

Nelson SM, Lawlor DA. Predicting live birth, preterm delivery, and low birth weight in infants born from in vitro fertilisation: A prospective study of 144,018 treatment cycles. *PLoS Med* 2011;**8**:e1000386.

Oakley L, Doyle P, Maconochie N. Lifetime prevalence of infertility and infertility treatment in the UK: results from a population-based survey of reproduction. *Hum Reprod* 2008;**23**:447–450.

Olivius C, Friden B, Borg G, Bergh C. Why do couples discontinue in vitro fertilization treatment? A cohort study. *Fertil Steril* 2004;**81**:258–61.

Oudshoorn SC, van Tilborg TC, Eijkemans MJC, Oosterhuis GJE, Friederich J, van Hooff MHA, van Santbrink EJP, Brinkhuis EA, Smeenk JMJ, Kwee J, *et al.* Individualized versus standard FSH dosing in women starting IVF/ICSI: an RCT. Part 2: The predicted hyper responder. *Hum Reprod* 2017;**32**:2506–2514.

Pandian Z, Marjoribanks J, Ozturk O, Serour G, Bhattacharya S. Number of embryos for transfer following in vitro fertilisation or intra-cytoplasmic sperm injection. *Cochrane database Syst Rev* 2013;**7**:CD003416.

Rajkhowa M, Mcconnell A, Thomas GE. Reasons for discontinuation of IVF treatment: a questionnaire study. *Hum Reprod* 2006;**21**:358–363.

Rossi B V, Berry KF, Hornstein MD, Cramer DW, Ehrlich S, Missmer SA. Effect of Alcohol Consumption on In Vitro Fertilization. *Obstet Gynecol* 2011;**117**:136–142.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. In: John Wiley & Sons, 2004.

Smeenk JM, Stolwijk AM, Kremer JA, Braat DD. External validation of the templeton model for predicting success after IVF. *Hum Reprod* 2000;**15**:1065–8.

Smith ADAC, Tilling K, Lawlor DA, Nelson SM. External Validation and Calibration of IVFpredict: A National Prospective Cohort Study of 130,960 In Vitro Fertilisation Cycles. Sun Q-Y (ed). *PLoS One* 2015;**10**:e0121357.

Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;**338**:b2393.

Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer New York, 2009.

Stockwell T, Zhao J, Greenfield T, Li J, Livingston M, Meng Y. Estimating under- and over-reporting of drinking in national surveys of alcohol consumption: identification of consistent biases across four English-speaking countries. *Addiction* 2016;**111**:1203–1213.

Templeton A, Morris JK, Parslow W. Factors that affect outcome of in-vitro fertilisation treatment. *Lancet* 1996;**348**:1402–1406.

te Velde ER, Nieboer D, Lintsen AM, Braat DDM, Eijkemans MJC, Habbema JDF, Vergouwe Y. Comparison of two models predicting IVF success; the effect of time trends on model performance. *Hum Reprod* 2014;**29**:57–64.

van Buuren S, Groothuis-Oudshoorn K. MICE : Multivariate Imputation by Chained Equations in R. *J Stat Softw* 2011;**45**:1–67.

van Loendersloot L, Repping S, Bossuyt PMM, van der Veen F, van Wely M. Prediction models in in vitro fertilization; where are we? A mini review. *J Adv Res* 2014;**5**:295–301.

van Loendersloot LL, van Wely M, Repping S, van der Veen F, Bossuyt PMM. Templeton prediction model underestimates IVF success in an external validation. *Reprod Biomed Online* 2011;**22**:597–602.

van Tilborg TC, Eijkemans MJ, Laven JS, Koks CA, de Bruin JP, Scheffer GJ, van Golde RJ, Fleischer K, Hoek A, Nap AW, *et al.* The OPTIMIST study: optimisation of cost effectiveness through individualised FSH stimulation dosages for IVF treatment. A randomised controlled trial. *BMC*

Womens Health 2012;**12**:29.

van Tilborg TC, Oudshoorn SC, Eijkemans MJC, Mochtar MH, van Golde RJT, Hoek A, Kuchenbecker WKH, Fleischer K, de Bruin JP, Groen H, *et al.* Individualized FSH dosing based on ovarian reserve testing in women starting IVF/ICSI: a multicentre trial and cost-effectiveness analysis. *Hum Reprod* 2017;**32**:2485–2495.

van Tilborg TC, Torrance HL, Oudshoorn SC, Eijkemans MJC, Koks CAM, Verhoeve HR, Nap AW, Scheffer GJ, Manger AP, Schoot BC, *et al.* Individualized versus standard FSH dosing in women starting IVF/ICSI: an RCT. Part 1: The predicted poor responder. *Hum Reprod* 2017;**32**:2496–2505.

Waylen AL, Metwally M, Jones GL, Wilkinson AJ, Ledger WL. Effects of cigarette smoking upon clinical outcomes of assisted reproduction: a meta-analysis. *Hum Reprod Update* 2009;**15**:31–44.

Wong KM, Mastenbroek S, Repping S. Cryopreservation of human embryos and its contribution to in vitro fertilization success rates. *Fertil Steril* 2014;**102**:19–26.

Zegers-Hochschild F, Adamson GD, Dyer S, Racowsky C, de Mouzon J, Sokol R, Rienzi L, Sunde A, Schmidt L, Cooke ID, *et al.* The International Glossary on Infertility and Fertility Care, 2017†‡§. *Hum Reprod* 2017;**32**:1786–1801.

Figure legends

Figure 1: Flow chart presenting the numbers (%) of live birth, treatment continuation and discontinuation over six complete cycles in the OPTIMIST and HFEA databases (McLernon *et al.*, 2016).

Figure 2: Calibration plots showing the association between the calculated and observed cumulative live birth rates over 3 complete IVF/ICSI cycles in the OPTIMIST cohort for **a)** the *original pre-treatment model* as described by McLernon et al (McLernon *et al.*, 2016) **b)** *recalibrated pre-treatment model* with adjustment of the intercept.

Figure 3: Calibration plot showing the association between the calculated and observed cumulative live birth rates over 3 complete IVF/ICSI cycles in the OPTIMIST cohort for the *original post-treatment model* as described by McLernon (McLernon *et al.*, 2016).

Figure 4: Example of the *recalibrated pre-treatment model* predicting the cumulative probability of a live birth up to three complete ICSI cycles for a woman with primary infertility caused by a male factor, aged 30 or 40 years with an infertility duration of two or five years.

Figure 5: Example of the with AMH, AFC and body weight *updated pre-treatment model* predicting the cumulative probability of a live birth up to three complete ICSI cycles for a woman with two years of primary infertility caused by a male factor, aged 30 or 40 years, a total body weight of 70 kilograms, with an AMH of 2.0 or 0.5 ng/mL and an AFC of 15 or 7.

Figure 6: Example of the *post-treatment model* predicting the cumulative probability of a live birth up to three complete ICSI cycles for a woman with two years of primary infertility caused by a male factor, aged 30 or 40 years, with 5 or 10 oocytes retrieved, a cleavage stage single embryo transfer, with or without embryo cryopreservation.

Supplementary Figure 1. Plots showing the adjusted relation between the predictors included in the *updated McLernon pre-treatment model* and the probability of a live birth after IVF/ICSI treatment.

Predictor; linear predictor (XB) of the original pre-treatment model as described by McLernon (McLernon et al. 2016), Weight; female body weight in kg, AFC; antral follicle count (2-10mm), AMH; anti-Müllerian hormone (ng/mL)

Supplementary Figure 2. Plots showing the adjusted relation between the predictors in the *updated McLernon post-treatment model* and the probability of a live birth after IVF/ICSI treatment.

Predictor: linear predictor (XB) of the original post-treatment model as described by McLernon (McLernon et al 2016); AFC; antral follicle count (2-10mm), AMH; anti-Müllerian hormone (ng/mL)

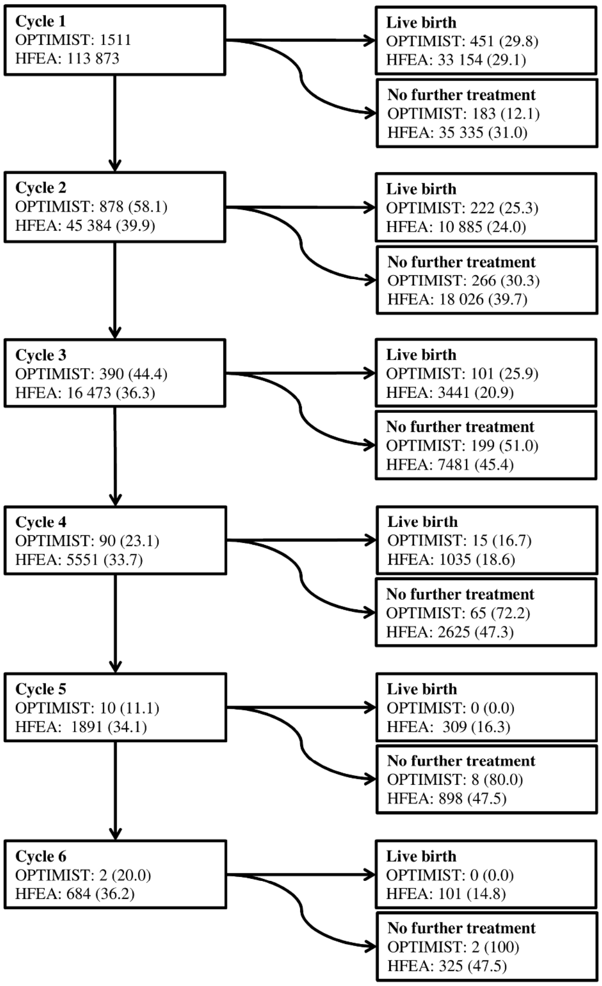
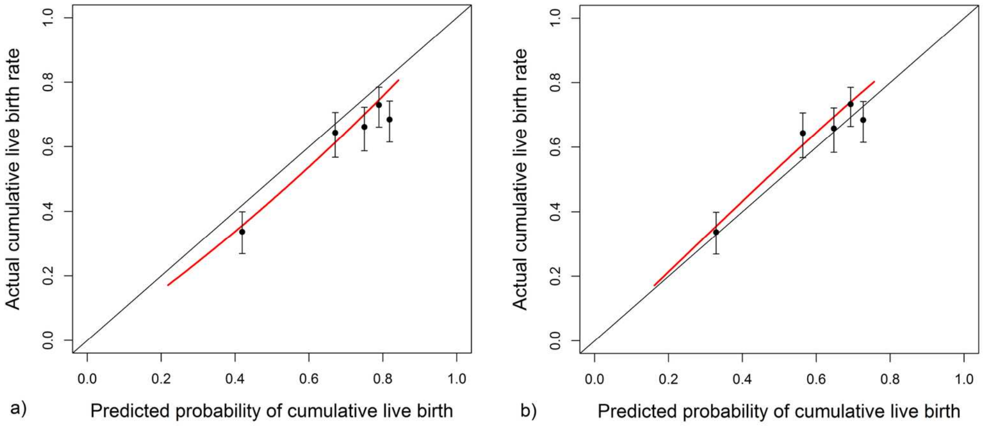
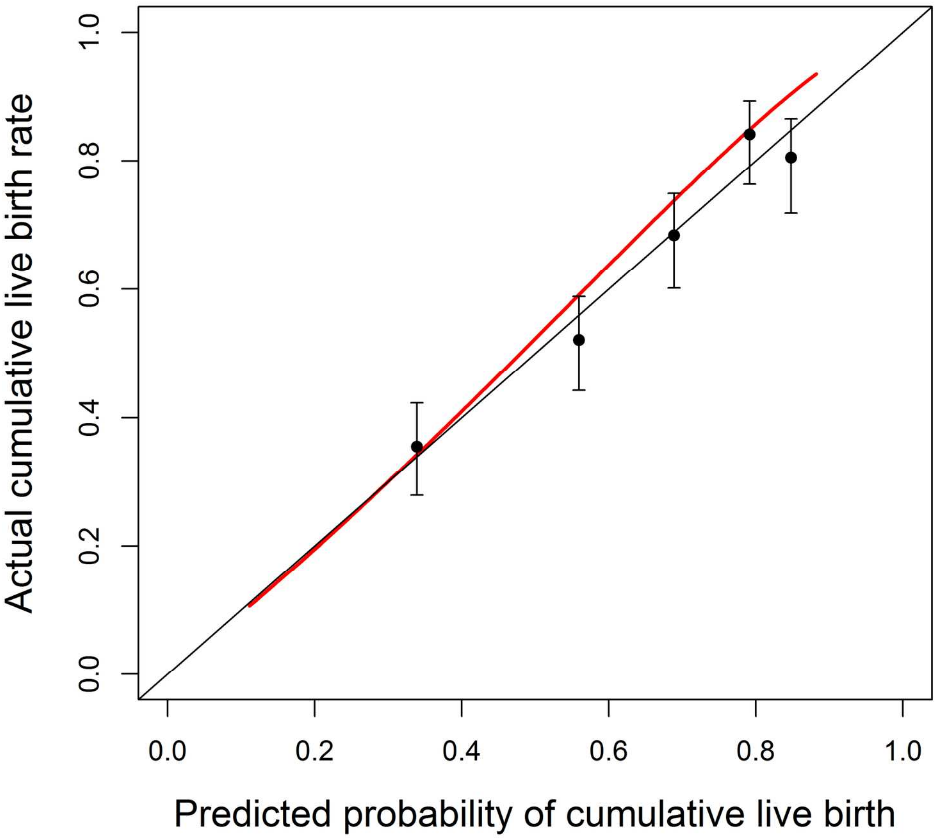


Figure 1: Flow chart presenting the numbers (%) of live birth, treatment continuation and discontinuation over six complete treatment cycles in the OPTIMIST and HFEA databases (McLernon et al., 2016).

209x297mm (300 x 300 DPI)



152x76mm (300 x 300 DPI)



152x152mm (300 x 300 DPI)

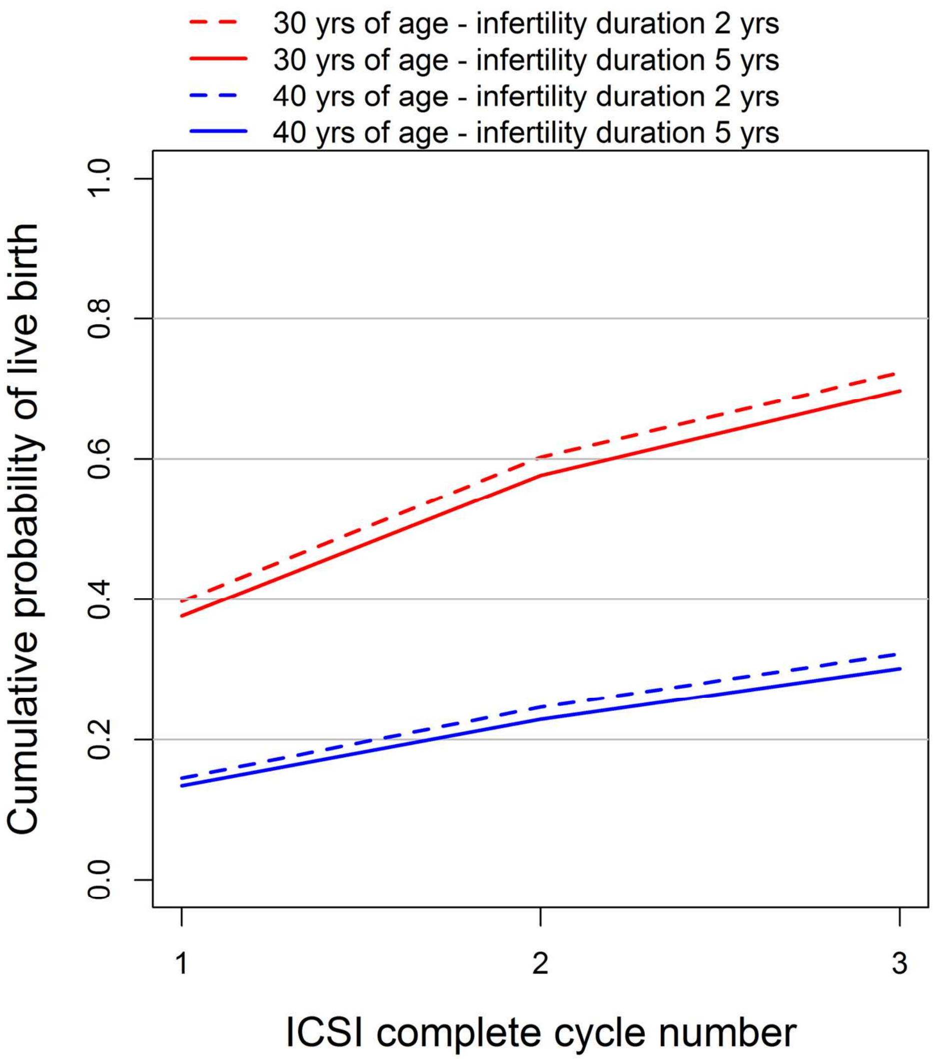


Figure 4: Example of the recalibrated pre-treatment model predicting the cumulative probability of live birth up to three complete ICSI cycles for a woman with primary infertility caused by a male factor, aged 30 or 40 years with an infertility duration of two or five years.

152x166mm (300 x 300 DPI)

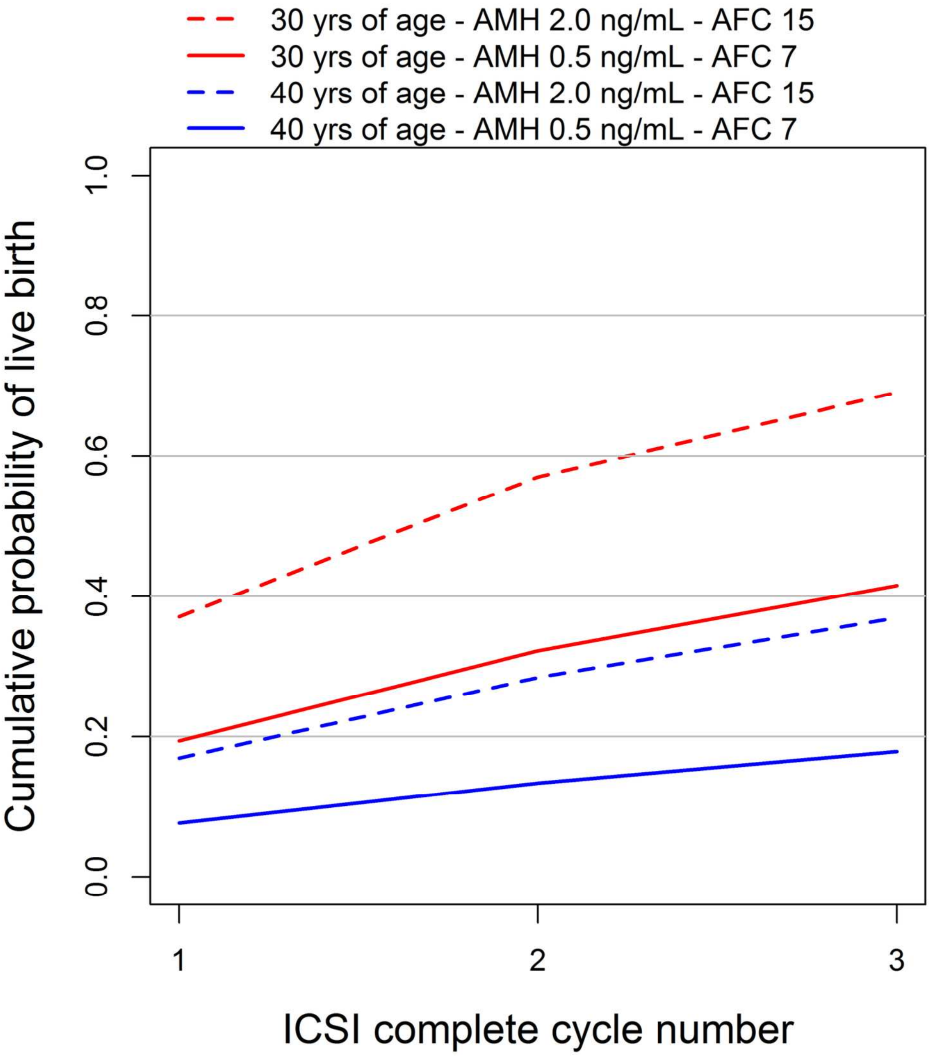
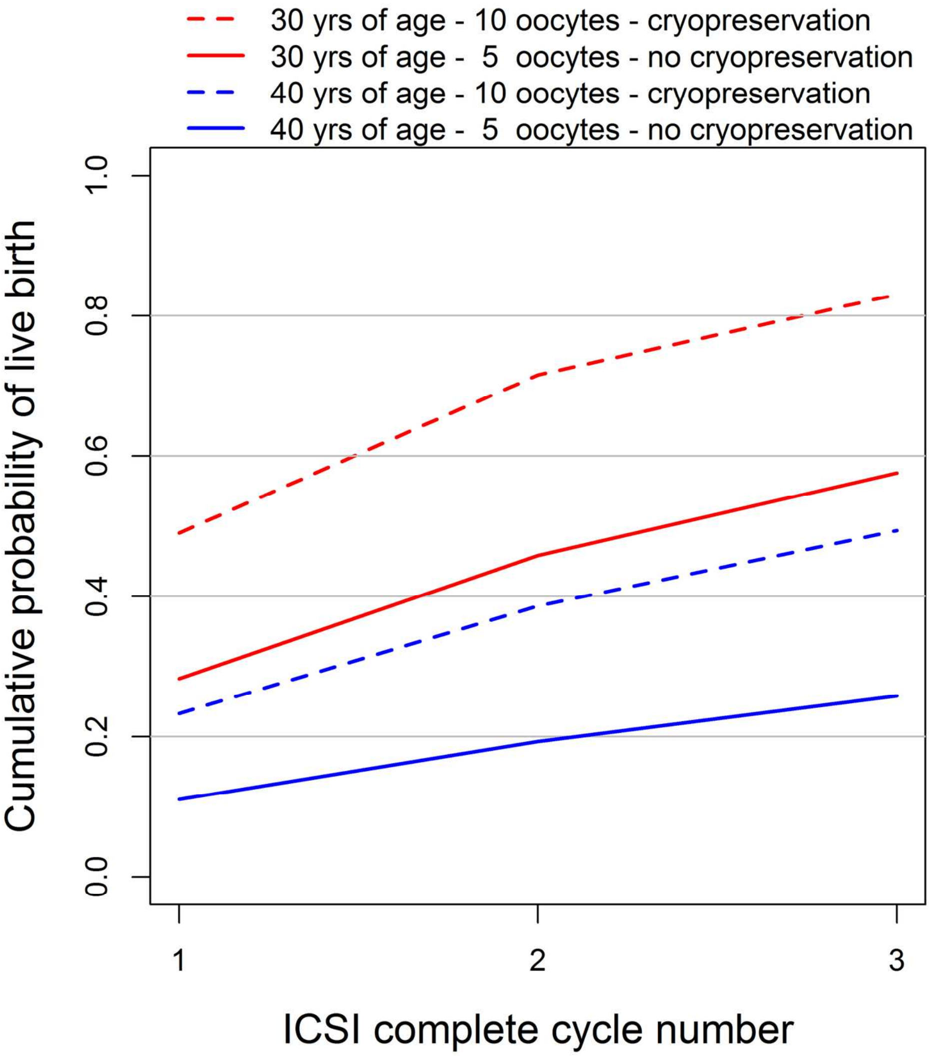


Figure 5: Example of the with AMH, AFC and body weight updated pre-treatment model predicting the cumulative probability of live birth up to three complete ICSI cycles for a woman with two years of primary infertility caused by a male factor, aged 30 or 40 years, a total body weight of 70 kilograms, with an AMH of 2.0 or 0.5ng/mL and an AFC of 15 or 7.

152x166mm (300 x 300 DPI)



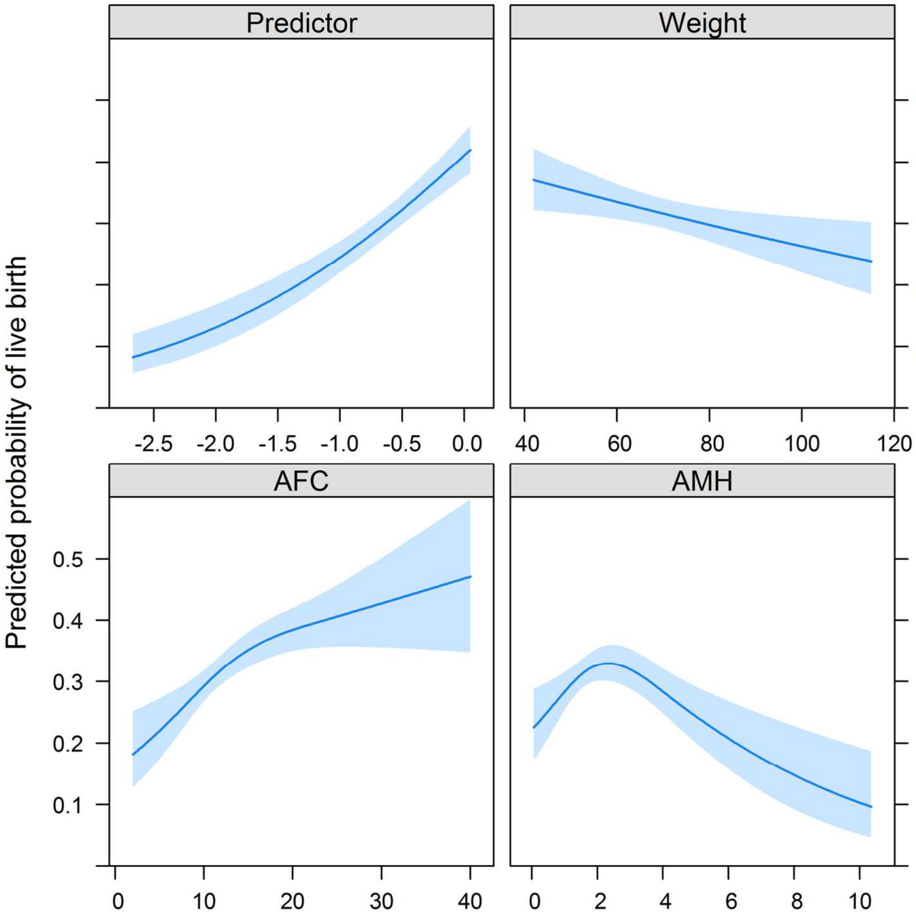
152x166mm (300 x 300 DPI)

Tables

Table I Characteristics of patient and treatment variables included as predictors in the development sample (HFEA cohort) and the validation sample (OPTIMIST cohort) (McLernon *et al.*, 2016). ~~Unless stated otherwise data are n (%).~~

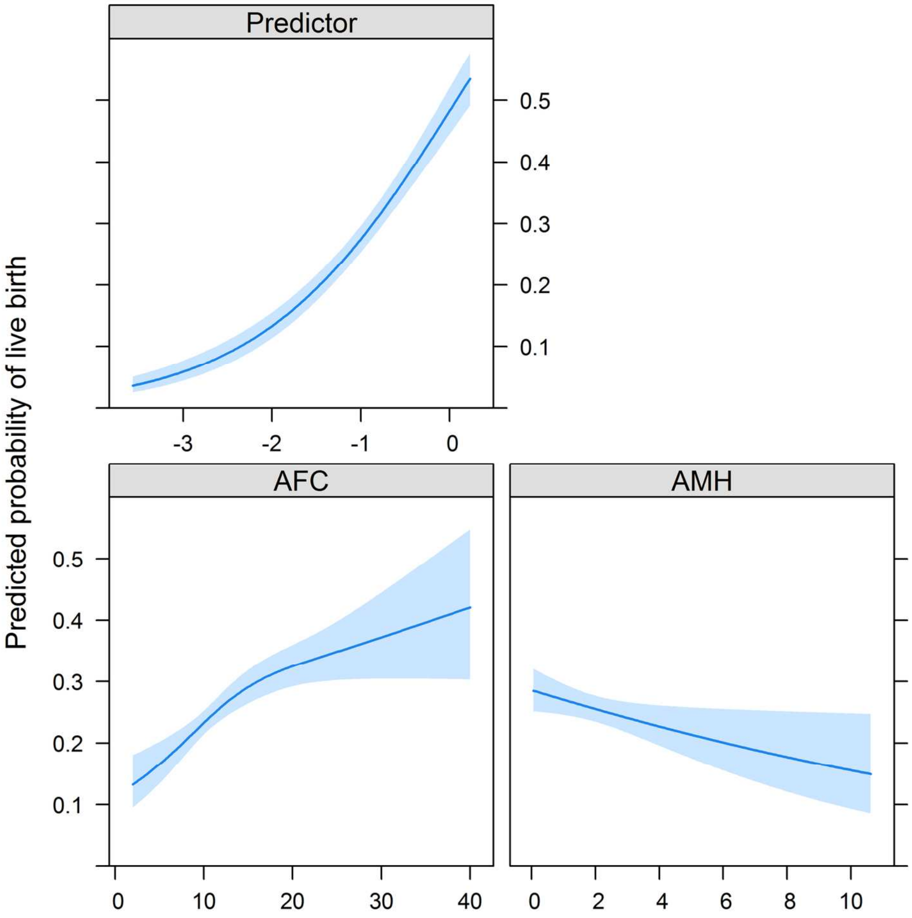
Characteristics	**HFEA cohort**	**OPTIMIST cohort**	**Missing values in OPTIMIST cohort (%)**
No of women	113 873	1 511	
No of complete cycles	184 269	2 881	
Patient characteristics			
Age (years), mean (SD)	34.1 (5)	33.5 (5)	2 (0.1)
Duration of infertility (years), median (IQR)	4 (3-6)	2 (2-3)	18 (1.2)
No previous pregnancy in couple~~,~~	75 541 (66)	917 (61)	2 (0.1)
Cause of infertility:			
- Tubal factor~~)~~	26 545 (23)	158 (11)	
- Male factor	49 753 (44)	839 (56)	
- Anovulatory	15 942 (14)	NA by protocol	
- Endometriosis	7 590 (7)	60 (4)	
- Unexplained	32 693 (29)	521 (35)	
Body weight (kg), mean (SD)	NA	69.5 (13)	36 (2.4)
Anti-Müllerian hormone (ng/mL), median (IQR)	NA	1.9 (1-3)	169 (11.2)
Antral follicle count (2-10mm), median (IQR)	NA	13 (9-18)	
Treatment characteristics of first completed cycle			
IVF	67 511 (59)	830 (55)	
ICSI	46 362 (41)	681 (45)	
No of oocytes retrieved, median (IQR)	8 (5-13)	8 (5-13)[a]	1 (0.1)
No of embryos created, median (IQR)	5 (2-8)	4 (2-7)[a]	4 (0.3)
No of embryos frozen, median (IQR)	0 (0-1)	1 (0-3)[a]	6 (0.5)
Cryopreservation of embryos	28 950 (25)	726 (48)	
Fresh embryo transfer: stage and no. of transferred embryos:			24 (1.6)
- Cleavage stage SET	9 248 (8)	1 004 (66)	
- Cleavage stage DET	75 701 (66)	125 (8)	
- Cleavage stage TET	8 649 (8)	4 (0.3)	
- Blastocyst stage SET	662 (1)	NA	
- Blastocyst stage DET	2 960 (3)	NA	
- Blastocyst stage TET~~)~~	130 (0.1)	NA	
- No transfer	15 501 (14)	354 (23)	

Data are presented as number (%) unless ~~Unless stated~~ otherwise specified. ~~data are n (%).~~ IQR; interquartile range, NA; not available, SET; single embryo transfer, DET; double embryo transfer, TET; triple embryo transfer. a) Median is calculated over 1293 women who had an ovarian follicle aspiration.



Supplementary Figure 1. Plots showing the adjusted relation between the predictors included in the updated McLernon pre-treatment model and the probability of a live birth after IVF/ICSI treatment. ‼ ┼ ‼ ┼ Predictor; linear predictor (XB) of the original pre-treatment model as described by McLernon (McLernon et al. 2016), Weight; female body weight in kg, AFC; antral follicle count (2-10mm), AMH; anti-Müllerian hormone (ng/mL).

152x152mm (300 x 300 DPI)



Supplementary Figure 2. Plots showing the adjusted relation between the predictors in the updated McLernon post-treatment model and the probability of a live birth after IVF/ICSI treatment.‼ ┼ ‼ ┼ Predictor: linear predictor (XB) of the original post-treatment model as described by McLernon (McLernon et al 2016); AFC; antral follicle count (2-10mm), AMH; anti-Müllerian hormone (ng/mL).

152x152mm (300 x 300 DPI)

Supplementary text 1. McLernon pre-treatment model.

Table showing the predictors in the original McLernon pre-treatment model (McLernon *et al.*, 2016)**.**

Name predictor	Description	Range of possible values
Age	Female age	18 to 50 years
Duration	How long have you been trying to conceive?	0 to 21 years
Previous	Have you been pregnant before?	1 = No; 0 = Yes
Tubal	Do you have a problem with your tubes?	1 = Yes; 0 = No
Anovulation	Do you have an ovulation problem?	1 = Yes; 0 = No
MaleFactor	Do you have a male factor fertility problem?	1 = Yes; 0 = No
Unexplained	Do you have an unexplained fertility problem?	1 = Yes; 0 = No
Treatment	Which fertility treatment are you planning on having?	1 = ICSI; 0 = IVF

Original pre-treatment model formulas as described by McLernon et al. (McLernon *et al.*, 2016)**.**

1. For the non-linear relation between Age and the probability of live birth, the following Age1, Age2 and Age3 equations are first calculated and then used in the XB equation below (point 3).

   - Age1 = max((Age-26)/k,0)**3+(11*max((Age-41)/k,0)**3-(15)*max((Age-37)/k,0)**3)/4;
   - Age2 = max((Age -31)/k,0)**3+(6*max((Age -41)/k,0)**3-(10)*max((Age -37)/k,0)**3)/4;
   - Age3 = max((Age -34)/k,0)**3+(3*max((Age -41)/k,0)**3-(7)*max((Age -37)/k,0)**3)/4; *k=15**(2/3); **means 'to the power of'*

2. For the non-linear relation between Year and the probability live birth, the following Year1 and Year2 equations are first calculated and then used in the XB equation below (point 3). The value Year= 0 is used for the most up to date predictions.

   - Year1 = max((Year+9)/k,0)**3+((6)*max((Year)/k,0)**3-(9)*max((Year+3)/k,0)**3)/(3);
   - Year2 = max((Year+6)/k,0)**3+((3)*max((Year)/k,0)**3-(6)*max((Year+3)/k,0)**3)/(3); *k= 9**(2/3).*

3. Calculate XB.

   XB = -0.9948 + 0.0362[a] + (0.0275*Age) + (-0.1805*Age1) + (0.4553*Age2) + (-1.1990*Age3) + (-0.0295*Duration) + (-0.0772*Previous) + (-0.0957*Tubal) + (0.0492*Anovulation) + (-0.1005*MaleFactor) + (0.0602*Unexplained) + (0.2155*Treatment) + (0.0334*Year) + (-0.0370*Year1) + (0.2173*Year2).

   a) To inflate predictions to 2013 an additional 0.0362 is added.

4. Calculate the predicted probability of live-birth after the first, second, …., sixth IVF cycle.

   PCycle1 = exp(XB)/(1+exp(XB))
   PCycle2 = exp(XB - 0.2394)/(1+exp(XB - 0.2394))
   PCycle3 = exp(XB - 0.4110)/(1+exp(XB - 0.4110))
   PCycle4 = exp(XB - 0.5628)/(1+exp(XB - 0.5628))

PCycle5 = exp(XB - 0.7189)/(1+exp(XB - 0.7189))
PCycle6 = exp(XB - 0.8138)/(1+exp(XB - 0.8138))

5. Calculate the predicted *cumulative* probability of a live-birth after 1, 2, 3,…., 6 completed IVF cycles:

CumPCycle1 = 1-(1-p1)
CumPCycle2 = 1-((1-p1)*(1-p2))
CumPCycle3 = 1-((1-p1)*(1-p2)*(1-p3))
CumPCycle4 = 1-((1-p1)*(1-p2)*(1-p3)*(1-p4))
CumPCycle5 = 1-((1-p1)*(1-p2)*(1-p3)*(1-p4)*(1-p5))
CumPCycle6 = 1-((1-p1)*(1-p2)*(1-p3)*(1-p4)*(1-p5)*(1-p6))

Supplementary text 2. McLernon post-treatment model.

Table showing the predictors in the original McLernon post-treatment model (McLernon *et al.*, 2016)**.**

Name predictor	Description	Range of possible values
Age	Female age	18 to 50 years
Duration	How long have you been trying to conceive?	0 to 21 years
Previous	Have you been pregnant before?	1 = No; 0 = Yes
Tubal	Do you have a problem with your tubes?	1 = Yes; 0 = No
Eggs	How many eggs were collected on your first IVF cycle?	(1 to 28)
Treat	Was your first cycle IVF or ICSI?	(1 = ICSI; 0 = IVF)
Cryo	In your first cycle did you have embryos frozen?	(1 = Yes; 0 = No)

Stage	What type of embryo transfer did you have in your first fresh embryo transfer?	(No embryos transferred;
		Single cleavage stage;
		Single blastocyst stage;
		Double cleavage stage;
		Double blastocyst stage;
		Triple cleavage stage;
		Triple blastocyst stage)

Original post-treatment model formulas as described by McLernon et al. (McLernon *et al.*, 2016)**:**

1. For the non-linear relation between Age and the probability of live birth, the following Age1, Age2 and Age3 equations are first calculated and then used in the XB equation below (point 4):

    - Age1 = max((Age-26)/k,0)**3+(11*max((Age-41)/k,0)**3-(15)*max((Age-37)/k,0)**3)/4;
    - Age2 = max((Age -31)/k,0)**3+(6*max((Age -41)/k,0)**3-(10)*max((Age -37)/k,0)**3)/4;
    - Age3 = max((Age -34)/k,0)**3+(3*max((Age -41)/k,0)**3-(7)*max((Age -37)/k,0)**3)/4;
      k=15**(2/3), **means 'to the power of'

2. For the non-linear relation between Year and the probability of live birth, the following Year1 equation is first calculated and then used in the XB equation below (point 4). The value Year = 0 is used for the most up to date predictions.

    - Year1 = max((Year+8)/k,0)**3+((4)*max((Year+1)/k,0)**3-(7)*max((Year+4)/k,0)**3)/(3);
      k= 7**(2/3).

3. For the non-linear relation between Eggs and the probability of live birth, the following Eggs1 equation is first calculated and then used in the XB equation below (point 4):

    - Eggs1=max((Eggs-3)/k,0)**3+((6)*max((Eggs-18)/k,0)**3-(15)*max((Eggs-9)/k,0)**3)/(9);
      k= 15**(2/3).

4. Calculate XB

XB = -1.7564 + 0.0362[a] + (0.0272*Age) + (-0.1556*Age1) + (0.3812*Age2) + (-1.0184*Age3) + (-0.0208*Duration) + (-0.0504*Previous) + (-0.2207*Tubal) + (0.0018*Year) +

(0.0619*Year1) + (0.0630*Eggs) + (-0.0479*Eggs1) + (-0.0968*Treat) + (0.6490*Cryo) + Stage[b]

a) To inflate predictions to 2013 an additional 0.0362 is added.
b) Stage equals the following values depending on group chosen:
If Double cleavage stage then Stage=0;
If No embryos transferred then Stage= -1.0842;
If Single cleavage stage then Stage= -0.5675;
If Single blastocyst stage then Stage= 0.0684;
If Double blastocyst stage then Stage= 0.5802;
If Triple cleavage stage then Stage= 0.0218;
If Triple blastocyst stage then Stage= 0.4547.

1. Calculate the predicted probability of live-birth after the first, second, …., sixth IVF cycle:

   PCycle1 = exp(XB)/(1+exp(XB))
   PCycle2 = exp(XB - 0.1933)/(1+exp(XB - 0.1933))
   PCycle3 = exp(XB - 0.3537)/(1+exp(XB - 0.3537))
   PCycle4 = exp(XB - 0.5122)/(1+exp(XB - 0.5122))
   PCycle5 = exp(XB - 0.6788)/(1+exp(XB - 0.6788))
   PCycle6 = exp(XB - 0.7666)/(1+exp(XB - 0.7666))

2. Calculate the predicted *cumulative* probability of a live-birth after 1, 2, 3,…., 6 complete IVF cycles:

   CumPCycle1 = 1-(1-p1)
   CumPCycle2 = 1-((1-p1)*(1-p2))
   CumPCycle3 = 1-((1-p1)*(1-p2)*(1-p3))
   CumPCycle4 = 1-((1-p1)*(1-p2)*(1-p3)*(1-p4))
   CumPCycle5 = 1-((1-p1)*(1-p2)*(1-p3)*(1-p4)*(1-p5))
   CumPCycle6 = 1-((1-p1)*(1-p2)*(1-p3)*(1-p4)*(1-p5)*(1-p6))

Supplementary Text 3. Recalibrated pre-treatment model.

Recalibrated pre-treatment model formula

The included predictors and formulas 1, 2, 4 and 5 are unchanged to the original McLernon pre-treatment model (see Supplementary Text 1)

3. Calculate XB.

XB = - 0.3474[a] - 0.9948 + 0.0362[b] + (0.0275*Age) + (-0.1805*Age1) + (0.4553*Age2) + (-1.1990*Age3) + (-0.0295*Duration) + (-0.0772*Previous) + (-0.0957*Tubal) + (0.0492*Anovulation) + (-0.1005*MaleFactor) + (0.0602*Unexplained) + (0.2155*Treatment) + (0.0334*Year) + (-0.0370*Year1) + (0.2173*Year2).

a) To recalibrate the pre-treatment model, 0.3474 is subtracted from the intercept.
b) To inflate predictions to 2013 an additional 0.0362 is added.