Human Reproduction, Vol.31, No.1 pp. 84–92, 2016
Advanced Access publication on October 25, 2015 doi:10.1093/humrep/dev268

human reproduction

ORIGINAL ARTICLE *Infertility*

Predicting the chance of live birth for women undergoing IVF: a novel pretreatment counselling tool

R.K. Dhillon[1,*], D.J. McLernon[2], P.P. Smith[1], S. Fishel[3], K. Dowell[3], J.J. Deeks[4], S. Bhattacharya[1], and A. Coomarasamy[1]

[1]School of Clinical and Experimental Medicine, University of Birmingham, Academic department, Birmingham Women's Hospital, Birmingham B15 2TG, UK [2]Division of Applied Health Sciences, School of Medicine and Dentistry, Foresterhill, Aberdeen AB25 2ZD, UK [3]CARE (Centres for Assisted Reproduction) John Webster House, 6 Lawrence Drive, Nottingham Business Park, Nottingham NG8 6PZ, UK [4]School of Health and Population Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

*Correspondence address. School of Clinical and Experimental Medicine, University of Birmingham, Academic department, Birmingham Women's Hospital, Birmingham B15 2TG, UK. E-mail: rima.dhillon@doctors.org.uk

Submitted on May 10, 2015; resubmitted on September 3, 2015; accepted on October 5, 2015

STUDY QUESTION: Which pretreatment patient variables have an effect on live birth rates following assisted conception?

SUMMARY ANSWER: The predictors in the final multivariate logistic regression model found to be significantly associated with reduced chances of IVF/ICSI success were increasing age (particularly above 36 years), tubal factor infertility, unexplained infertility and Asian or Black ethnicity.

WHAT IS KNOWN ALREADY: The two most widely recognized prediction models for live birth following IVF were developed on data from 1991 to 2007; pre-dating significant changes in clinical practice. These existing IVF outcome prediction models do not incorporate key pretreatment predictors, such as BMI, ethnicity and ovarian reserve, which are readily available now.

STUDY DESIGN, SIZE, DURATION: In this cohort study a model to predict live birth was derived using data collected from 9915 women who underwent IVF/ICSI treatment at any CARE (*Centres for Assisted Reproduction*) clinic from 2008 to 2012. Model validation was performed on data collected from 2723 women who underwent treatment in 2013. The primary outcome for the model was live birth, which was defined as any birth event in which at least one baby was born alive and survived for more than 1 month.

PARTICIPANTS/MATERIALS, SETTING, METHODS: Data were collected from 12 fertility clinics within the CARE consortium in the UK. Multivariable logistic regression was used to develop the model. Discriminatory ability was assessed using the area under receiver operating characteristic (AUROC) curve, and calibration was assessed using calibration-in-the-large and the calibration slope test.

MAIN RESULTS AND THE ROLE OF CHANCE: The predictors in the final model were female age, BMI, ethnicity, antral follicle count (AFC), previous live birth, previous miscarriage, cause and duration of infertility. Upon assessing predictive ability, the AUROC curve for the final model and validation cohort was (0.62; 95% confidence interval (CI) 0.61–0.63) and (0.62; 95% CI 0.60–0.64) respectively. Calibration-in-the-large showed a systematic over-estimation of the predicted probability of live birth (Intercept (95% CI) = −0.168 (−0.252 to −0.084), $P < 0.001$). However, the calibration slope test was not significant (slope (95% CI) = 1.129 (0.893–1.365), $P = 0.28$). Due to the calibration-in-the-large test being significant we recalibrated the final model. The recalibrated model showed a much-improved calibration.

LIMITATIONS, REASONS FOR CAUTION: Our model is unable to account for factors such as smoking and alcohol that can affect IVF/ICSI outcome and is somewhat restricted to representing the ethnic distribution and outcomes for the UK population only. We were unable to account for socioeconomic status and it may be that by having 75% of the population paying privately for their treatment, the results cannot be generalized to people of all socioeconomic backgrounds. In addition, patients and clinicians should understand this model is designed for use *before* treatment begins and does not include variables that become available (oocyte, embryo and endometrial) as treatment progresses. Finally, this model is also limited to use prior to first cycle only.

WIDER IMPLICATIONS OF THE FINDINGS: To our knowledge, this is the first study to present a novel, up-to-date model encompassing three readily available prognostic factors; female BMI, ovarian reserve and ethnicity, which have not previously been used in prediction models for IVF outcome. Following geographical validation, the model can be used to build a user-friendly interface to aid decision-making for couples and their clinicians. Thereafter, a feasibility study of its implementation could focus on patient acceptability and quality of decision-making.

STUDY FUNDING/COMPETING INTEREST: None.

Key words: prediction model / live birth / IVF / assisted conception / counselling

Introduction

The number of couples seeking IVF in the UK has continued to rise, with a 3% increase from the 62 158 couples treated in 2012 to 64 000 in 2013 ('Fertility treatment in 2013. Trends and figures. HFEA.'). Contrary to common perception, IVF does not guarantee success; between 38 and 49% of couples who start IVF will remain childless, even after undergoing up to six IVF cycles (Malizia *et al.*, 2009). It is therefore important that subfertile couples are well informed about their chances of success with IVF. Based on their specific probability of success, the couple can decide whether the risks of the treatment and the emotional and, in many cases, financial burden can be justified. To optimize counselling for couples on their chances of a live birth after IVF, clinical prediction models, which estimate the chance of an outcome adjusted for a patient's characteristics, may play a role since clinicians' judgments can often be inaccurate (Wiegerinck *et al.*, 1999; Van Der Steeg *et al.*, 2006). Reliance on annually published validated age-stratified national success rates (Australian Government Department of Health and Ageing, 2006; Andersen *et al.*, 2009) has meant that clinicians often tend to base predictions solely on age.

There have been many attempts to build prediction models to aid clinicians in predicting IVF success (Stolwijk *et al.*, 1996, 1998; Templeton *et al.*, 1996; Minaretzis *et al.*, 1998; Hunault *et al.*, 2002, 2007; Nelson and Lawlor, 2011). The two most widely recognized models, which used live birth as the primary outcome, are those by Templeton *et al.* (1996) and Nelson and Lawlor (2011). A study by a Dutch team (te Velde *et al.*, 2014) used their cohort to validate both these models to assess the effects of time trends on model performance. They found that the Templeton model underestimated success rates, as one may expect given that it is a much older study, and the Nelson model overestimated success rates. The study showed that the calibration of both models considerably improved when the models were adjusted for the changing success rates over time.

A recent study by Smith *et al.* also performed external validation of the Templeton and Nelson models using a large dataset of over 130 000 cycles (Smith *et al.*, 2015). They found that the discriminative power (assessed using area under receiver operating characteristic (AUROC) curve) was comparable between the models; but that the Nelson model had markedly better calibration. They also found both models underestimated the live birth rate, although as seen with te Velde *et al.* (2014), this improved when the models were updated to reflect improvements in live birth rates over time.

A recent report by the Human Fertilisation and Embryology Authority (HFEA) recognized that IVF practice and outcomes have seen significant changes between 2008 and 2012, primarily because of the introduction of Day 5 (blastocyst) embryo transfer ('Fertility treatment in 2013. Trends and figures. HFEA.'). Given these advancements in technology, and the fact that most existing models were developed before 2008, there is a need for a new model based on more recent data.

Another pitfall of the existing models is their inability to account for certain key predictors of IVF treatment outcome. In particular, the most recent of these models (Nelson and Lawlor, 2011) built using a large dataset provided by the HFEA was not able to include BMI, any measure of ovarian reserve or ethnicity. A systematic review in 2011 which included 33 studies concluded that a raised BMI has an adverse effect on pregnancy outcomes for women underdoing IVF treatment (Rittenberg *et al.*, 2011). They also found that this negative association was apparent for both obese (BMI 30–39.9) and overweight (BMI 25–29.9) women (Rittenberg *et al.*, 2011). There is also strong evidence to suggest that women with a diminished ovarian reserve generally have a poor response to gonadotrophin therapy and therefore the chance of a successful pregnancy (Ulug *et al.*, 2003; Jirge, 2011). A recent study reported that antral follicle count (AFC) correlated strongly with the number of mature oocytes retrieved in IVF/ICSI cycles (Shaban and Abdel Moety, 2014), which, in turn, can influence the chances of pregnancy. Another study found that AFC provided additional prognostic value to female age in predicting response to ovarian hyperstimulation (Broer *et al.*, 2013). Finally, several large cohort studies have shown that ethnicity has an association with IVF outcome, with Black and South Asian women appearing to have the poorest outcomes (Seifer *et al.*, 2008, 2010; Baker *et al.*, 2010; Fujimoto *et al.*, 2010; Luke *et al.*, 2011). In addition, a review published in 2012 concluded that current evidence suggests there are significant disparities in IVF outcomes between ethnic groups (Wellons *et al.*, 2012). Despite this, ethnicity is a factor that is yet to be included as a predictor in any model predicting live birth following IVF.

The aim of this study is to derive, assess and validate a novel predictive model that will estimate the chance of live birth for women undergoing their first IVF non-donor cycle. This model will use only pretreatment factors and include previously unrecorded predictors such as BMI, ovarian reserve and ethnicity.

Materials and Methods

Derivation cohort

The study population was derived from a database of all patients who had undergone their first fresh non-donor cycle of IVF (including ICSI) at any of the *Centres for Assisted Reproduction* (CARE) clinics across the UK and Ireland, between 2008 and 2012. CARE is one of the UK's largest independent providers of fertility services, where both NHS and non-NHS patients are treated, ~25% of patients are NHS funded and 75% fund themselves. The CARE database consists of routinely collected baseline demographics, cycle data and outcome data for all patients.

Within the variable for previous IVF, any woman with a history of IVF treatment, whether it was at a CARE clinic or elsewhere, was assigned a '1', women without any history of IVF treatment were assigned '0'. All women with a '1' were excluded from analysis. The reason for this was to exclude previous treatment as a confounder and also because the primary use of the model is for couples seen at their first clinic appointment, *prior* to embarking on IVF treatment. The decision to include IVF and ICSI as one variable was because the authors agreed that success rates are comparable for the two

treatment modalities and so it was reasonable to include them together. Also, the model is designed for use before patients undergo treatment, and because occasionally in the cases of 'mild male factor' clinicians will often decide to crossover from IVF to ICSI once the patient has come through for treatment we felt it was better to keep IVF and ICSI as one variable.

Baseline demographics, cycle data and outcome data were retrieved from 12 CARE clinics across the UK. The CARE consortium is composed of five main fertility clinics (Nottingham, Manchester, Northampton, Sheffield and Dublin) and a further seven satellite centres. For patients seen initially at the satellite clinics, they are seen up to the point of egg collection; egg collection, all embryology and embryo transfer are then performed at the nearest main clinic. Following the embryo transfer the satellite clinic resumes full care of the patient.

The original database contained information on over 50 000 cycles dating back to 1998. A decision was made to limit the dataset from 2008 onwards due to advances in technology over time and improvements in clinical practice, such as greater numbers of blastocyst transfer and single embryo transfer, as detailed in the recent HFEA report, which in turn have affected success rates ('Fertility treatment in 2013. Trends and figures. HFEA.'). Data from the first cycle only were used to eliminate the bias from previous cycle failures. Furthermore, by limiting to only first cycle we were able to express the probability of live birth outcome per individual woman.

Primary outcome

The primary outcome for the model was live birth, which was defined as any birth event in which at least one baby was born alive and survived for more than 1 month. This definition is consistent with previous publications, including that of Templeton *et al.* (1996) and Nelson and Lawlor (2011).

Statistical analyses

Model development

Univariable logistic regression analyses were performed to assess the association of each of the predictive factors with live birth. A multivariable logistic regression model was used to derive the final prediction model for live birth. The predictors included in the multivariable model were pre-selected based on knowledge from the existing literature (van Loendersloot *et al.*, 2010) and clinical knowledge and were as follows: age, BMI, ethnicity, cause of infertility, duration of infertility, AFC, previous live birth and previous miscarriage. AFC was selected in preference to early follicular FSH as it is a more accurate measure of ovarian reserve (Jirge, 2011; Broer *et al.*, 2013). Anti-mullerian hormone (AMH) has similar accuracy to AFC (Jirge, 2011; Broer *et al.*, 2013) and is a more objective measure of ovarian reserve. However, as AMH is a fairly recent test it was not available for most patients in the derivation cohort, therefore AFC was selected in preference.

The continuous variables of age, BMI, duration of infertility and AFC were assessed for their functional form using plots of the observed log odds (Supplementary Figs S1–S4). In the case of age, duration of infertility and AFC there was a non-linear relationship with live birth. Appropriate transformations were carried out and subsequently included in the model. The results for age showed that below 36 years of age the chances of live birth appeared fairly constant, but above 36 there was a sharp linear decline, resulting in two linear variables being created for age.

Missing data

The whole dataset contained 9915 women, data entry was complete in all variables except for BMI and AFC, therefore we were required to impute the missing data. A multiple imputation procedure was conducted using an iterative Markov chain Monte Carlo method. All predictors and the outcome of live birth were included in the imputation process to maximize the precision of the imputations. All univariable models and the multivariable model were fitted to the 20 imputed datasets arising from the multiple imputation procedure. The parameter estimates and covariance's arising from the models from each imputed dataset were combined to produce inferential results.

Predictive ability

Initially the model was assessed for predictive ability using apparent validation. Apparent validation is when the model performance is assessed directly in the same cohort from which it was derived (Steyerberg, 2009). The two performance measures used were discrimination and calibration. Discrimination is the ability of the model to correctly discriminate between those who had the outcome and those that did not i.e. correctly distinguish between the women who had a live birth (for whom the model assigns a higher probability) and women who do not have a live birth (for whom the model assigns a lower probability). The AUROC curve (also known as a c-statistic) was used as a measure of discrimination. Calibration refers to the agreement between the predicted probabilities of live birth and the observed (actual) probabilities. The predicted probabilities from the final model were assessed for accuracy across increasing tenths of predicted probabilities using calibration plots. The mean observed probability is plotted against the mean predicted probability in each tenth and perfect calibration is displayed as a straight line passing through zero with a gradient of one.

Model validation

External validation of the model was performed on a cohort of women undergoing their first fresh IVF cycle at any CARE clinic during the year of 2013 (temporal validation) (Steyerberg, 2009). The missing data in the validation cohort were also imputed using the same method as the derivation cohort. For ease of computation and interpretation, the average measures of the imputed values were taken across all 20 imputed datasets for women who had values imputed, so that validation was performed on only one dataset. The model was fitted to the validation cohort (2013 population) using the same parameter estimates derived from the study cohort (2008–2012 population). The predictive ability of the model was assessed on the external validation cohort. The AUROC curve was determined to assess discriminatory ability and calibration plots were presented. As a formal test of calibration we assessed calibration-in-the-large to compare the mean predicted probability of live birth with the mean observed probability of live birth. This is essentially the intercept from the model, which is only adjusted for the linear predictors (as an offset) from the final model, applied to the patients in the external cohort. A significant deviation from zero indicates that predictions are systematically too low or too high (Steyerberg and Vergouwe, 2014). The calibration slope was also calculated, where a perfect slope (i.e. perfect agreement between predicted and observed probabilities) would have a gradient of one. Significant deviations from one would suggest that low predicted probabilities were too low or too high and high predicted probabilities were too high or too low.

All statistical analyses were performed using the Statistical Package for the Social Sciences (ver. 21.0; SPSS Inc., Chicago, IL, USA) and SAS (ver.9.2; SAS Institute, Cary, NC, USA). A value of $P < 0.05$ was considered significant.

Ethical approval

Permission for use of the database was granted by the CARE IRB following review of the study protocol. The dataset was anonymized according to the ICO's (Information Commissioner's Office) guide on non-identifiable data. Furthermore the CARE data protection certificate allows for their data to be used for survey and research purposes.

Data sharing

The complete anonymized dataset is held on a secure, password protected University of Birmingham account belonging to the corresponding author

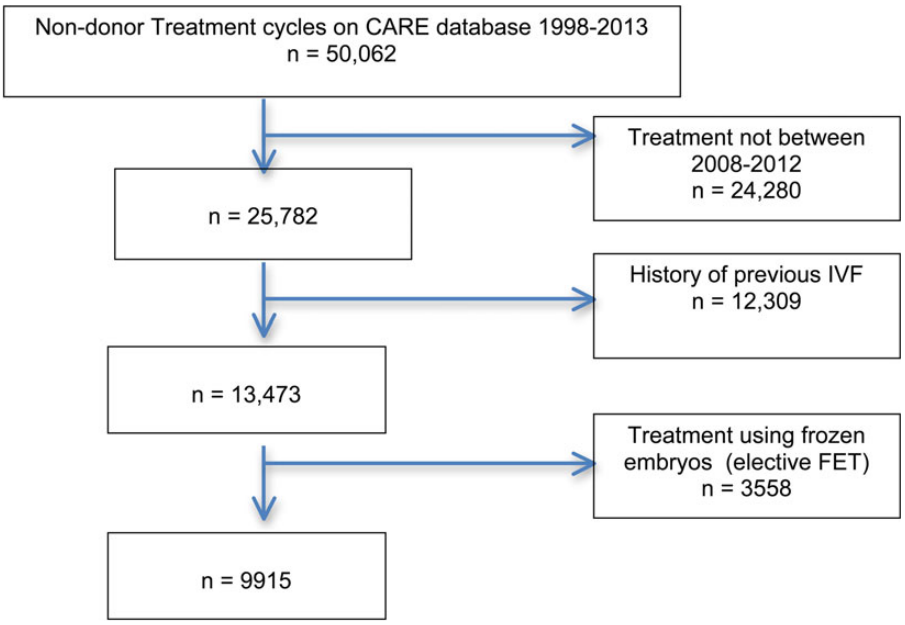

Figure 1 Definition of eligible cohort and analysis samples. The flowchart displays the process of selection for the study population and the corresponding number of cycles. Total number of cycles used for analysis, $n = 9915$.

only. Participant consent was not obtained as the presented data are anonymized and risk of identification is extremely low.

Results

Derivation cohort

A total of 9915 women were used to build the final model. Figure 1 shows how we established the eligible cohort of IVF (including ICSI) treatment cycles. Table I shows the baseline characteristics of the cohort. The overall rate of at least one live birth from the whole dataset was 31.5%.

Missing data

Only two of the variables selected for use in the multivariate model had missing data, these were BMI and AFC. Descriptive characteristics of women with missing and non-missing data for BMI and AFC can be found in Supplementary Table SI. The data across each baseline characteristic were reasonably comparable between the two groups. However, significantly more women with a BMI measurement were of white ethnicity (81.7 versus 70.5%, $P < 0.001$) and had partners with male factor infertility (65.1 versus 48.5%, $P < 0.001$) than women without a BMI measurement.

Univariate and multivariate analyses

The univariate associations of the potential predictors for live birth following IVF are shown in Table II. The multivariable logistic regression model predicting live birth is displayed in Table III. The model shows that the odds of a successful live birth decrease with age. This reduction in the odds of live birth is greater with each increasing year of age past the age of 36 years compared with up to the age of 36 years. Other variables which showed a statistically significant reduction in odds of live birth in the multivariate final model were; tubal factor, unexplained infertility, and

Table I Baseline characteristics of the cohort undergoing a first IVF (including ICSI) treatment cycle.

	Cohort *n* (%) or mean (SD)
	Whole dataset (*n* = 9915)
Age (years)	34.6 (5.4)
Duration of infertility (in completed years)	2.0 (2.0)
BMI*	24.8 (4.0)
AFC*	18.7 (13.6)
Previous miscarriage	1818 (18.3%)
Previous live birth	1578 (15.9%)
Cause of infertility	
Tubal factor	1442 (14.5%)
Anovulation	1088 (11.0%)
Unexplained	2950 (29.8%)
Other (e.g. endometriosis, fibroids)	3005 (30.3%)
Male factor	5611 (56.6%)
Ethnicity	
White	7530 (75.9%)
Asian	768 (7.7%)
Black	162 (1.6%)
Chinese	60 (0.6%)
Other	115 (1.2%)
Not stated	924 (9.3%)
Mixed	356 (3.6%)

*Variable contains missing data.
AFC, antral follicle count.

Table II Univariate associations of potential predictors for live birth following IVF.

	OR (95% CI)	*P*-value
Age (years)	0.94 (0.93–0.94)	<0.001
Duration of infertility (in completed years)	0.97 (0.95–0.99)	0.003
BMI*	0.98 (0.97–0.99)	0.01
AFC*	1.01 (1.01–1.02)	<0.001
Previous miscarriage	0.84 (0.74–0.94)	0.002
Previous live birth	0.93 (0.83–1.05)	0.2
Cause of infertility		
Tubal factor	0.90 (0.80–1.01)	0.08
Anovulation	1.21 (1.07–1.40)	0.003
Unexplained	1.01 (0.92–1.11)	0.8
Other (e.g. Endometriosis, fibroids)	0.70 (0.63–0.77)	<0.001
Male factor	1.02 (0.94–1.11)	0.7
Ethnicity		
White	Reference	
Asian	0.95 (0.81–1.12)	0.6
Black	0.44 (0.29–0.67)	<0.001
Chinese	0.84 (0.48–1.48)	0.5
Other	0.68 (0.44–1.05)	0.08
Not stated	1.00 (0.86–1.16)	1
Mixed	0.86 (0.68–1.09)	0.2

*Variable contains missing data.
OR, odds ratio; CI, confidence interval.

being Asian or Black. The univariate analysis suggested that increasing BMI, duration of infertility >5 years and previous miscarriage were associated with decreased odds of live birth, while increasing AFC was associated with a significantly increased odds of live birth. However these associations became non-significant in the multivariate analysis.

Predictive ability

The AUROC curve test for discriminatory ability of the final prediction model for odds of live birth was 0.62 (95% confidence interval (CI) 0.61–0.63). The ROC curve and calibration plots are displayed in Supplementary Fig. S5a and b, respectively.

Model validation

Our external cohort consisted of 2723 patients who had undergone their first fresh assisted treatment cycle at any CARE clinic in the year of 2013. The baseline characteristics, cycle characteristics and outcome data for the validation cohort are displayed in Supplementary Table SII. The overall live birth rate for this cohort was 31.7%. The baseline characteristics of the both the derivation and validation cohorts were comparable, as were the overall live birth rates. The AUROC for the final model applied to the external cohort was 0.62 (95% CI 0.60 to 0.64). Calibration-in-the-large showed a systematic over-estimation of the predicted probability of live birth (Intercept (95% CI) = −0.168 (−0.252 to −0.084), $P < 0.001$). However, the calibration slope test was not significant (slope (95% CI) = 1.129 (0.893 to 1.365), $P = 0.28$) meaning that the over-estimation was uniform across the range of predicted probabilities (Fig. 2). Due to the calibration-in-the-large test being significant we recalibrated the final model. This was done by scaling the linear predictor from the final model, using the slope and intercept ($y = -0.078 + 1.129$); we then adjusted for the final model linear predictor and applied this to the external cohort. The recalibrated model is shown in Fig. 3 and shows a much improved calibration.

Discussion

To date, successful prediction of live birth after assisted reproductive technology has been limited. We have developed a novel model, which encompasses prognostic factors that have not previously been used, such as BMI, ovarian reserve and ethnicity. The key predictors in our model that were shown to have a significant effect on the chances of live birth are: age, tubal factor, unexplained causes of infertility and being South Asian or Black.

Strengths and weaknesses

This is the first successfully derived and externally validated prediction model for live birth following assisted conception for women undergoing their first fresh non-donor cycle of treatment, which accounts for BMI, ethnicity and ovarian reserve. This prediction model is purposefully simple, in that its use is only for women undergoing their first fresh non-donor cycle. We believe this prediction tool holds an important role as an adjunct in the counselling process for women at the critical decision-making point in their journey, i.e. *before* they embark on their first treatment cycle. The advantage of using data from a first IVF cycle means that the calculated probabilities are expressed per woman/couple and not per cycle.

The greatest strength of our model is that it has highlighted ethnicity as a key predictor for IVF success; ethnicity is a factor which has not been used in any previous prediction models. Ethnicity has been recognized in many American papers (Seifer *et al.*, 2008, 2010; Baker *et al.*, 2010; Fujimoto *et al.*, 2010; Luke *et al.*, 2011; Wellons *et al.*, 2012) as a confounding factor in affecting IVF success and we have seen this also in our model. There appears to be a strong association between being South Asian or Black and having a lower chance of live birth even when accounting for the other predictors in the multivariate analysis. We recently published a cohort study and meta-analysis investigating the effect of ethnicity on IVF outcome (Dhillon *et al.*, 2015) and explored potential reasons why South Asian and Black women have lower live birth rates following IVF, including accounting for fibroids in Black women, however no solid explanations could be drawn. However, despite the addition of ethnicity as a novel key predictor to our model, given the large variation in ethnic groups across the globe, our model is somewhat restricted to representing the ethnic distribution and outcomes for the UK population only. It would be useful to externally validate this model on a dataset from a different country to see if ethnic variability affects the performance measures of the model. A further limitation of the inclusion of ethnicity within the model is that the group with 'not stated' ethnicity constitutes more than 10% of the study population, in addition all the ethnic minority groups are smaller than this 'not stated' group and so this may have influenced the data and added bias to the results.

In addition to ethnicity, no previous models have accounted for BMI or AFC. As mentioned in the results, the univariate analysis for BMI and live birth outcome was statistically significant, showing that increasing BMI

Table III Final multivariate logistic regression model for live birth (n = 9915).

	Parameter estimate	SE	*P*-value	Odds ratio	95% CI	
					Lower	Upper
Age:						
≤36 years	−0.035589	0.008	<0.001	0.97	0.95	0.98
>36 years	−0.106139	0.012	<0.001	0.90	0.88	0.92
BMI	−0.010881	0.009	0.2	0.99	0.97	1.01
Cause of infertility:						
Male factor	−0.085967	0.056	0.1	0.91	0.82	1.02
Tubal factor	−0.254369	0.069	<0.001	0.78	0.68	0.89
Anovulation	−0.138708	0.082	0.09	0.87	0.74	1.02
Unexplained	−0.133782	0.067	0.04	0.88	0.77	0.99
Other (e.g. endometriosis)	−0.118451	0.062	0.05	0.89	0.79	1.00
Ethnicity:						
White	0			Reference		
Asian	−0.171572	0.084	0.04	0.84	0.71	0.99
Black	−0.683648	0.214	<0.001	0.51	0.33	0.77
Chinese	−0.181580	0.293	0.5	0.83	0.47	1.48
Other	−0.355212	0.222	0.1	0.70	0.45	1.08
Not stated	−0.005533	0.083	0.9	0.99	0.84	1.17
Mixed	−0.192857	0.122	0.1	0.83	0.65	1.05
Previous live birth						
No	0			Reference		
Yes	0.093953	0.063	0.1	1.10	0.97	1.24
Previous miscarriage						
No	0			Reference		
Yes	−0.023788	0.060	0.7	0.98	0.87	1.10
AFC	0.015095	0.008	0.06	1.02	1.00	1.03
AFC (squared)	−0.000142	0.000	0.2	1.00	1.00	1.00
Duration of infertility:						
0–4 years	0			Reference		
≥5 years	−0.093313	0.066	0.2	0.91	0.80	1.04
Constant	0.811547	0.355	0.02	2.25	1.12	4.54

reduces the odds of live birth, however this association became non-significant in the final model. This could be explained by the fact that other predictors in the model carry more weight in influencing live birth when looked at in combination: one of the strongest predictors, as we would expect, was female age. It appears from our data that BMI increases with increasing age and so this would explain why in multivariate analysis, where age is accounted for, the effect of BMI on live birth is not significant. In addition to this, it appears that in general Black women have higher BMI than White women. Black ethnicity alone is a strong predictor for lower chances of IVF success; after accounting for ethnicity this could be another reason why the association between BMI and reduced IVF success is lost in the multivariate analysis.

Similarly as for BMI, the univariate analysis for AFC and live birth was significant, showing that increasing AFC is associated with a higher odds of live birth, however this became non-significant in the final model. When looking at the data further, it shows that women with increasing age have reduced AFC and so, as was seen with BMI, the effect of AFC on IVF outcome after accounting for age is reduced; nonetheless on this basis alone we feel AFC should not be rejected as a predictor in the model. Furthermore, we acknowledge that AFC is a subjective measure and therefore open to intra-observer variability, however it has been shown that even with this variability, its ability to predict IVF success is comparable with AMH (a non-subjective measure of ovarian reserve) (Bonilla-Musoles *et al.*, 2012; Tremellen and Savulescu, 2014). Furthermore, recording of AMH was very poor within the database and therefore, in order to use a variable with fewer missing entries, AFC was selected over AMH.

Inevitably, in any prediction model, one is unable to account for the residual confounding effect of the unavailable variables. One of the weaknesses of our model is that we have been unable to account for confounders such as smoking status and alcohol intake. A recent systematic review and meta-analysis on predictive factors in IVF evaluated nine predictive factors: female age, duration of subfertility, type of subfertility, indication for IVF, basal FSH, fertilization method, number of oocytes,

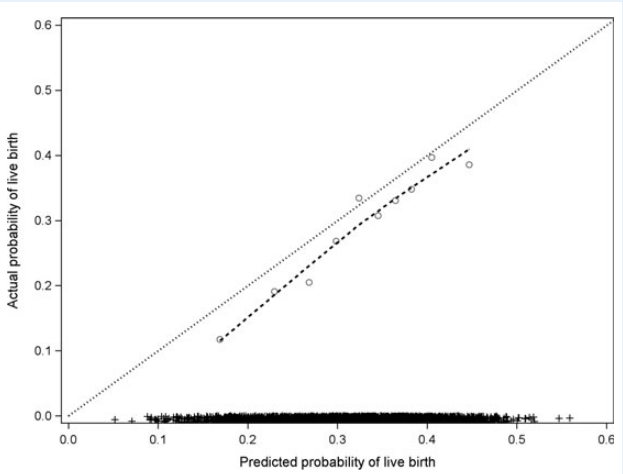

Figure 2 Graphical representation of the predicted probability of live birth against the actual probability of live birth. Patients were ranked into order of predicted probability of live birth and divided into tenths. The circles represent the mean risks for each tenth; the dotted line represents the perfect relationship; the dashed line represents the smooth non-parametric Loess calibration curve fitted through the circles; the plus symbols represent the spread of patients across predicted risks.

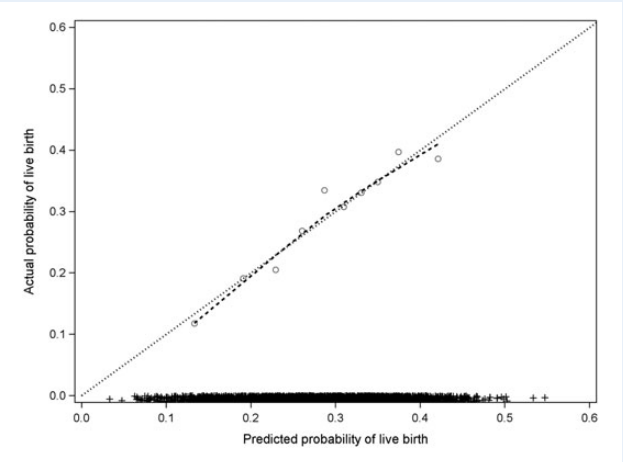

Figure 3 Graphical representation of the predicted probability of live birth against the actual probability of live birth following recalibration. Recalibration of data from Fig. 2 was done by scaling the linear predictor from the final model, using the slope and intercept ($y = -0.078 + 1.129$); we then adjusted for the final model linear predictor and applied this to the external cohort ($n = 2723$ cycles).

number of embryos transferred, and embryo quality (van Loendersloot *et al.*, 2010). As our model is for pretreatment counselling only, we did not include any oocyte or embryo factors. We have, however, accounted for the other mentioned factors with the exception of basal FSH, where instead we have used a more accurate ovarian reserve measure in AFC.

Given the complexities of assisted conception there are many other confounders that can have an effect at different time points. For example there are prognostic factors which are only determined once a cycle has begun, such as oocyte number and embryo quality. A further limitation of our model is that it is restricted to use prior to starting treatment only. We appreciate that IVF success rates depend on more than the factors in this model alone. Therefore it is important for clinicians when using the model to ensure their patients understand that their probability of having a successful outcome will invariably change as they progress through their treatment and thus should be interpreted as a baseline prediction only.

A final weakness is that this model was built on a predominantly self-funded population. As we have been unable to account for socioeconomic status it may be that by having 75% of the population paying privately for their treatment this cannot be generalized to people of all socioeconomic backgrounds.

Comparison to existing models

Using our novel model, one is able to predict the chances of live birth following IVF, and this predictive ability has been assessed by the AUROC curve. Our model is the first model to incorporate new predictors such as ethnicity, AFC and BMI; with ethnicity shown to be a strong predictor of success. There are some similarities between our model and the existing models. The Templeton model gives the possibility of the category 'no previous IVF' and has been externally validated, and the validated Nelson model also gives the possibility of the categories 'no previous IVF' with 0, at least 1 pregnancy and at least 1 live birth. Thus both of these models can be applied before IVF is started and predict the success of the 1st IVF cycle, as for ours. Given the similarities in the use of our model compared with the existing models we have provided a crude comparison. Following apparent validation, Templeton *et al.* (1996) found the AUROC curve to be 0.62 (95% CI 0.61–0.62) and Nelson and Lawlor (2011) 0.63 (95% CI 0.62–0.64), whilst our model showed an AUROC curve of 0.62 (95% CI 0.61–0.63). Following external validation of our study, the AUROC curve was 0.62 (95% CI 0.60–0.64), the recently externally validated Nelson model (IVFpredict) and Templeton model had an AUROC of 0.63 (0.62–0.63) and 0.62 (0.61–0.62) respectively (Smith *et al.*, 2015), showing that our model has comparable discriminatory ability with these previous models. The Dutch study (te Velde *et al.*, 2014) and the more recent study by Smith *et al.* (2015) showed improvements in the performance of the Nelson and Templeton models when taking into account the effect of time trends. However, for our model there was no significant difference in live birth rates between 2008 and 2013 ($P = 0.2$). Adding treatment year to our model made no difference in the performance (AUROC 0.62, 95% CI 0.61–0.62) and so it was not included. A likely explanation is that both the Templeton and Nelson models were built on considerably older datasets compared with our model, pre-dating significant changes in clinical practice that occurred from 2008 onwards, therefore requiring an adjustment for time.

For IVF prediction models, calibration is deemed to be a more important measure of predictive ability than discrimination. A systematic review by Coppus *et al.* concluded that prediction models in reproductive medicine will be limited to an AUROC of no greater than 0.65 due to the relatively homogeneous group of subfertile patients (Coppus *et al.*, 2009). The calibration assessments for our model showed that there was a small systematic over-estimation in the predicted probabilities. After recalibration to correct for this, the calibration plot was much improved.

Table IV Examples of the final model to calculate predicted probability of live birth.

Example couples	Predicted probability of live birth	Recalibrated predicted probability
A. A 38-year-old White woman and her partner have been trying to conceive for over 5 years. She has a BMI of 35 kg/m^2 and an AFC of 14. The couple had a miscarriage in the past following a natural conception. The couple's cause of infertility is male factor infertility.	0.25	0.21
B. If we take the same couple as in A but change the BMI to 25 kg/m^2	0.27	0.24
C. If we take the same couple as in A but change the age to 30 years and ethnicity to Black.	0.21	0.18
D. A 28-year-old White woman with unexplained fertility. She has a BMI of 22 kg/m^2 and AFC of 15. The couple have had a child after a previous natural conception and have been trying for 2 years.	0.43	0.41

Note: These examples are plausible in terms of the types of patients that are seen in IVF clinics, and they show the influence of couple characteristics. A woman with a normal BMI (Example B) has a greater chance of success compared with a woman with a raised BMI (Example A) when keeping all other characteristics the same; this shows there is some influence of BMI on IVF outcome. Example D shows the characteristics that result in a greater chance of success.

Clinical implications

Examples of how our novel prediction model could be used in clinical practice to give an estimate of a couple's probability of achieving a live birth are shown in Table IV; an example of how to use the model is presented in Supplementary Fig. S6. We have presented the predicted probabilities for both the original externally validated model and the recalibrated model; the results show that the predicted probabilities from the recalibrated model were only slightly lower than those from the original model. We have illustrated not only the clinical use of this model (which will be developed into a freely available computer program, and/or mobile phone application) but also how a couple's characteristics influence their prognosis. This model provides a personalized approach to counselling and estimates chances of success based on easily measurable variables that are specific to the individual woman; rather than using success rates based on age-related national HFEA data. The idea would be for clinicians to use the model routinely when counselling couples seen in outpatient clinics for the first time, as the vast majority of UK hospital clinics will have computers with internet access. This should ensure that all patients have the opportunity to use the model at some point, which is particularly important for those patients who may have limited access to the internet or a mobile phone.

Future research

The next step for our model will be to further validate by performing geographical external validation. We plan to do this using the data collected from the Birmingham Women's Hospital Fertility Centre, as well as other assisted conception units. Following this we intend to build the model into a user-friendly web-based decision aid and mobile application allowing for use by both clinicians and patients. Finally, we intend to study patient experience of the tool and its impact on decision-making.

Supplementary data

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

Acknowledgements

We thank CARE for providing us with the database for the study.

Authors' roles

R.K.D., D.J.M, S.B. and A.C. generated the initial research idea. R.K.D., D.J.M. and P.P.S. performed all statistical analyses. R.K.D. wrote the manuscript. All authors reviewed and edited the final manuscript.

Funding

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

Conflict of interest

None declared.

References

Andersen AN, Goossens V, Bhattacharya S, Ferraretti AP, Kupka MS, de Mouzon J, Nygren KG. Assisted reproductive technology and intrauterine inseminations in Europe, 2005: results generated from European registers by ESHRE: ESHRE. The European IVF Monitoring Programme (EIM), for the European Society of Human Reproduction and Embryology (ESHRE). *Hum Reprod* 2009;**24**:1267–1287.

Australian Government Department of Health and Ageing. Report of the independent review of assisted reproductive technologies. Available: http://www.health.gov.au/internet/main/publishing.nsf/Content/ART-Report. 2006. (September 2014, n.d., date last accessed).

Baker VL, Luke B, Brown MB, Alvero R, Frattarelli JL, Usadi R, Grainger DA, Armstrong AY. Multivariate analysis of factors affecting probability of pregnancy and live birth with in vitro fertilization: an analysis of the Society for Assisted Reproductive Technology Clinic Outcomes Reporting System. *Fertil Steril* 2010;**94**:1410–1416. doi:10.1016/j.fertnstert.2009.07.986.

Bonilla-Musoles F, Castillo JC, Caballero O, Pérez-Panades J, Bonilla F, Dolz M, Osborne N. Predicting ovarian reserve and reproductive outcome using antimüllerian hormone (AMH) and antral follicle count (AFC) in patients with previous assisted reproduction technique (ART) failure. *Clin Exp Obstet Gynecol* 2012;**39**:13–18.

Broer SL, Dólleman M, van Disseldorp J, Broeze KA, Opmeer BC, Bossuyt PMM, Eijkemans MJC, Mol BW, Broekmans FJM, IPD-EXPORT Study Group. Prediction of an excessive response in in vitro fertilization from patient characteristics and ovarian reserve tests and comparison in

subgroups: an individual patient data meta-analysis. *Fertil Steril* 2013; **100**:420–429.e7. doi:10.1016/j.fertnstert.2013.04.024.

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

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. doi:10.1016/j.rbmo.2015.05.015.

Fertility treatment in 2013. Trends and figures. HFEA, n.d.

Fujimoto VY, Luke B, Brown MB, Jain T, Armstrong A, Grainger DA, Hornstein MD, Society for Assisted Reproductive Technology Writing Group. Racial and ethnic disparities in assisted reproductive technology outcomes in the United States. *Fertil Steril* 2010;**93**:382–390. doi:10.1016/j.fertnstert.2008.10.061.

Hunault CC, Eijkemans MJC, Pieters MHEC, te Velde ER, Habbema JDF, Fauser BCJM, Macklon NS. A prediction model for selecting patients undergoing in vitro fertilization for elective single embryo transfer. *Fertil Steril* 2002;**77**:725–732.

Hunault CC, te Velde ER, Weima SM, Macklon NS, Eijkemans MJC, Klinkert ER, Habbema JDF. A case study of the applicability of a prediction model for the selection of patients undergoing in vitro fertilization for single embryo transfer in another center. *Fertil Steril* 2007;**87**:1314–1321. doi:10.1016/j.fertnstert.2006.11.052.

Jirge PR. Ovarian reserve tests. *J Hum Reprod Sci* 2011;**4**:108–113. doi:10.4103/0974-1208.92283.

Luke B, Brown MB, Stern JE, Missmer SA, Fujimoto VY, Leach R. Racial and ethnic disparities in assisted reproductive technology pregnancy and live birth rates within body mass index categories. *Fertil Steril* 2011; **95**:1661–1666. doi:10.1016/j.fertnstert.2010.12.035.

Malizia BA, Hacker MR, Penzias AS. Cumulative live-birth rates after in vitro fertilization. *N Engl J Med* 2009;**360**:236–243. doi:10.1056/NEJMoa0803072.

Minaretzis D, Harris D, Alper MM, Mortola JF, Berger MJ, Power D. Multivariate analysis of factors predictive of successful live births in in vitro fertilization (IVF) suggests strategies to improve IVF outcome. *J Assist Reprod Genet* 1998;**15**:365–371.

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. doi:10.1371/journal.pmed.1000386.

Rittenberg V, Seshadri S, Sunkara SK, Sobaleva S, Oteng-Ntim E, El-Toukhy T. Effect of body mass index on IVF treatment outcome: an updated systematic review and meta-analysis. *Reprod Biomed Online* 2011; **23**:421–439. doi:10.1016/j.rbmo.2011.06.018.

Seifer DB, Frazier LM, Grainger DA. Disparity in assisted reproductive technologies outcomes in black women compared with white women. *Fertil Steril* 2008;**90**:1701–1710. doi:10.1016/j.fertnstert.2007.08.024.

Seifer DB, Zackula R, Grainger DA, Society for Assisted Reproductive Technology Writing Group Report. Trends of racial disparities in assisted reproductive technology outcomes in black women compared with white women: Society for Assisted Reproductive Technology 1999 and 2000 vs. 2004–2006. *Fertil Steril* 2010;**93**:626–635. doi:10.1016/j.fertnstert.2009.02.084.

Shaban MM, Abdel Moety GAF. Role of ultrasonographic markers of ovarian reserve in prediction of IVF and ICSI outcome. *Gynecol Endocrinol* 2014; **30**:290–293. doi:10.3109/09513590.2013.875996.

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. *fPLoS One* 2015;**10**:e0121357. doi:10.1371/journal.pone.0121357.

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

Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;**35**:1925–1931.

Stolwijk AM, Zielhuis GA, Hamilton CJ, Straatman H, Hollanders JM, Goverde HJ, van Dop PA, Verbeek AL. Prognostic models for the probability of achieving an ongoing pregnancy after in-vitro fertilization and the importance of testing their predictive value. *Hum Reprod* 1996; **11**:2298–2303.

Stolwijk AM, Straatman H, Zielhuis GA, Jansen CA, Braat DD, van Dop PA, Verbeek AL. External validation of prognostic models for ongoing pregnancy after in-vitro fertilization. *Hum Reprod* 1998;**13**:3542–3549.

Templeton A, Morris JK, Parslow W. Factors that affect outcome of in-vitro fertilisation treatment. *Lancet* 1996;**348**:1402–1406. doi:10.1016/S0140-6736(96)05291-9.

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.

Tremellen K, Savulescu J. Ovarian reserve screening: a scientific and ethical analysis. *Hum Reprod* 2014;**29**:2606–2614.

Ulug U, Ben-Shlomo I, Turan E, Erden HF, Akman MA, Bahceci M. Conception rates following assisted reproduction in poor responder patients: a retrospective study in 300 consecutive cycles. *Reprod Biomed Online* 2003;**6**:439–443. doi:10.1016/S1472-6483(10)62164-5.

Van Der Steeg J, Steures P, Eijkemans M, Habbema J, Bossuyt P, Hompes P, Van Der Veen F, Mol B. Do clinical prediction models improve concordance of treatment decisions in reproductive medicine? *BJOG* 2006;**113**:825–831. doi:10.1111/j.1471-0528.2006.00992.x.

van Loendersloot LL, van Wely M, Limpens J, Bossuyt PMM, Repping S, van der Veen F. Predictive factors in in vitro fertilization (IVF): a systematic review and meta-analysis. *Hum Reprod Update* 2010;**16**:577–589.

Wellons MF, Fujimoto VY, Baker VL, Barrington DS, Broomfield D, Catherino WH, Richard-Davis G, Ryan M, Thornton K, Armstrong AY. Race matters: a systematic review of racial/ethnic disparity in Society for Assisted Reproductive Technology reported outcomes. *Fertil Steril* 2012; **98**:406–409. doi:10.1016/j.fertnstert.2012.05.012.

Wiegerinck MAHM, Bongers MY, Mol BWJ, Heineman M-J. How concordant are the estimated rates of natural conception and in-vitro fertilization/embryo transfer success? *Hum Reprod* 1999;**14**:689–693.