Quality of clinical prediction models in in vitro fertilisation: which covariates are really important to predict cumulative live birth and which models are best?

David J McLernon[a], Siladitya Bhattacharya[b]

[a] Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK, AB25 2ZD, d.mclernon@abdn.ac.uk

[b] School of Medicine, Medical Sciences & Nutrition, University of Aberdeen, Aberdeen, UK, AB25 2ZD, s.bhattacharya@abdn.ac.uk

Corresponding author:

David J McLernon, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK, AB25 2ZD, d.mclernon@abdn.ac.uk

Abstract

The improvement in IVF cryopreservation techniques over the last 20 years has led to an increase in elective single embryo transfer, thus reducing multiple pregnancy rates. This strategy of successive transfers of fresh followed by frozen embryos has resulted in the acceptance of cumulative live birth over complete cycles of IVF as a critical measure of success. Clinical prediction models are a useful way of estimating the cumulative chances of success for couples tailored to their individual clinical factors which help them prepare for, and plan future treatment. In this review we describe several models that predict cumulative live birth and recommend which should be used by couples and/or their clinician and when they should be used. We also discuss the most relevant predictors to consider when either developing new IVF prediction models or updating existing models.

Key words

In-vitro fertilisation, clinical prediction models, predictors, ovarian reserve, female age.

Background

Over the last 15 years IVF practice has shifted from predominantly transferring multiple fresh embryos at a time to transferring a single fresh embryo (preferably a blastocyst) followed by successive episodes involving the transfer of single frozen-thawed embryos [1-3]. This change has been triggered by improvement in extended culture and embryo cryopreservation techniques. Such practice has seen the reduction of multiple pregnancies without compromising live birth rates and has led to a shift in the way outcomes are reported [4,5]. The traditional focus on live birth rates per fresh cycle has expanded to incorporate cumulative live birth rates which reflect the impact of frozen embryo replacements following an initial fresh transfer as well as subsequent treatment episodes [6-9]. Cumulative live birth rates are more helpful to couples and clinicians since they allow them to plan their care over a period of time [10]. While useful for getting an overall picture of IVF success at a national or clinic level, average cumulative live birth rates are not suitable for personalised medicine given that many patient and treatment level characteristics can affect chances of live birth in every couple [11]. A way of estimating the chance of live birth by factoring in all of these important characteristics is to use clinical prediction models.

Clinical prediction models are mathematical equations that allow us to combine a number of patient characteristics to predict an outcome in an individual [12]. These models can be used to predict the chance of a diagnosis or a consequence of a medical condition over a specified period of time. The former is usually termed a diagnostic model and the latter a prognostic model. In reproductive medicine we are usually concerned with predicting pregnancy outcomes by means of prognostic models. For prediction modelling we are primarily interested in the absolute risk of an individual given their personal characteristics. The term absolute risk refers to the chance that a patient will have the outcome over some specified time period e.g., a 20 year old women with unexplained infertility may have a 20% chance of a live birth without IVF over the next two years. Relative risk concerns the chance of the outcome occurring for one group of patients compared with some other group e.g., the

chance of live birth without treatment over the next two years for the average woman with endometriosis relative to the average woman with unexplained infertility may be a half. Since the term ‘risk’ is often used for unfavourable outcomes, we tend to use the term ‘chance’ for favourable outcomes such as live birth.

Clinical prediction models have different uses which must be decided before they are developed. They can be useful for providing evidence-based input for shared decision making around interventions such as choice of treatment, increased (or decreased) monitoring or referral to specialist care. They can also be useful to counsel patients or stratify patients by disease severity for treatment or research (e.g., inclusion in randomised trials). Specifically, IVF prediction models may be useful for informing patients of their individual chance of having a baby in order to manage their expectations and allow them to prepare physically, emotionally and (where relevant) financially for future treatment. In this review we examine existing IVF prediction models which attempt to predict the cumulative chances of live birth over several complete IVF cycles (i.e., cycles involving fresh and frozen embryos created from a single oocyte retrieval episode) using clinical data. We will provide recommendations as to which models are best for clinical and patient use. We will also consider which predictors are the most important to include for researchers wishing to validate or revise existing IVF models.

Cumulative live birth prediction models

In early 2020, a systematic review found 35 IVF prediction models in existence and reported on their methodological quality and predictive performance [13].

Three models (published before the end of January 2019) predicted cumulative live birth per woman [14,15]. The first estimated cumulative live birth up to three fresh embryo transfer attempts but excluded any subsequent frozen embryo transfers [14]. The other two models were developed using national UK data that predict the cumulative risk of live birth over

multiple complete cycles of IVF [15]. The term ‘complete cycle’ is used throughout this review and is always defined as all fresh and frozen embryo transfers resulting from a single episode of ovarian stimulation. The pre-treatment model calculates the cumulative chance of live birth and should only be used before starting the first IVF cycle. The predictors in this model include complete cycle number, female age, duration of infertility, previous pregnancy status, cause of infertility (tubal factor, male factor, anovulation or unexplained infertility) and type of treatment planned (IVF versus intracytoplasmic sperm injection (ICSI)). The post-treatment model revises the prediction at the time the woman undergoes her first embryo transfer and includes extra treatment specific predictors such as number of eggs collected, number of embryos transferred (0 to 3) and age of embryo i.e., blastocyst or cleavage stage. These models (available as the OPIS prediction calculators here: https://w3.abdn.ac.uk/clsm/opis) were also subsequently externally validated on an independent prospective cohort of 1515 women from The Netherlands [16] (see Table 1). The pre-treatment model had a relatively low c-statistic of 0.62 in the external cohort and needed recalibration. The c-statistic is a measure of model discrimination. To understand what discrimination means in this context, imagine a random pair of patients from the external cohort where one patient actually had a live birth and the other did not. The model should ideally have calculated a higher predicted chance of live birth for the patient who had the baby. If we repeat this for all possible pairs then the proportion correctly assigned a higher prediction gives us the c-statistic. A c-statistic of 0.5 means that our model is no better at distinguishing between low and high risk patients than a coin toss, while a c-statistic of 1 means the model is perfect (which is never the case). Calibration, on the other hand, is concerned with agreement between the predicted and the observed events and is ideally assessed using a flexible calibration curve [17]. The post-treatment model performed better with a c-statistic of 0.71 and did not require recalibration. The validation study also updated the models by adding BMI, anti-Müllerian hormone (AMH) and antral follicle count (AFC). All three improved the discrimination in the pre-treatment model (c-statistic=0.66) while no improvement was found in the post-treatment model (c-statistic=0.71). The post-treatment

model was adjusted for the number of eggs collected which could also be seen as a reflection of ovarian reserve. On the basis of these results, the additional value of the ovarian reserve tests can be questioned when a prediction model includes treatment information such as number of eggs, given the extra cost and physical burden associated with them. Female age is known to be correlated with ovarian reserve which may reduce the added value of these tests [18]. The post-treatment model was recommended in the review by Ratna et al (2021) [13], on the basis of its methodology, predictive performance and quality of reporting [13,15,19]. However, the pre-treatment model (which had lower discrimination) is arguably more useful, given that its intended moment of application is before IVF begins. The biggest limitation of these UK models is that the data used to develop the model are over 13 years old, which may affect the accuracy of the model when applied to today's patients. The HFEA data did not have some potentially important predictors which could have been included such as BMI, paternal age, alcohol intake, smoking and markers of ovarian reserve.

Models of note since the Ratna systematic review (2020 to 2022)

Since the Ratna review, two further model development studies are worthy of discussion. Both have predicted cumulative live birth over multiple complete cycles (Table 1).

USA

Two prediction models have been generated in one study which used national data from the Society for Assisted Reproductive Technology (SART) in the USA [20]. A pre-treatment model estimates the individualised chance of cumulative live birth over the first three complete cycles. The post-treatment model predicts chances before starting the second complete cycle in couples whose first complete cycle was unsuccessful. The model is available as a prediction tool at sart.org. The pre-treatment model was adjusted for female

age, previous full-term birth status, type of infertility (male factor, polycystic ovary syndrome, uterine factor, diminished ovarian reserve and unexplained infertility) and the female's BMI. A second pre-treatment model was also created for women who had an AMH measurement. Age, BMI and AMH had a non-linear relationship with live birth and so were included in the models as restricted cubic spline terms. As the value of AMH level increased so did the odds of live birth until around 5 ng/mL when it steadied. A woman with an AMH of 5 ng/mL had 22% increased odds of live birth compared to a woman with an AMH level of 2.5 ng/mL. Unfortunately, due to limitations of the dataset the authors could not include AMH as a predictor in the post-treatment model. They also could not assess the impact of clinics using different AMH assays, and although their performance was good in the SART data, the models have yet to be externally validated using independent datasets.

UK

A further UK based IVF prediction model was recently published by the same research group who developed the 2016 models and OPIS calculator [15,21]. When a couple have concluded their first complete cycle of IVF and have not achieved a live birth, they may decide to undergo a second complete cycle. Couples who were successful may decide to have more children. The previous UK models can only be used to estimate the chance of live birth either before commencing the first IVF treatment or at the first embryo transfer attempt which makes it more challenging for couples to prepare for the next stage of treatment. Using these models when the couple have finished their first complete cycle will not result in accurate predictions because they were developed using patient data measured before the first cycle. By the start of the second complete cycle, patient predictor values will have changed e.g., they will be older, the duration of infertility will be longer, and their cause of infertility may have changed. Further, many of the patients used to develop the models will not have had a second complete cycle which means that the case mix will have changed. Further prognosticators from the first complete cycle, such as the number of eggs collected

and the pregnancy outcome, will also be known. All of this new information was included in a model developed to estimate the chance of live birth in couples beginning a second complete cycle of IVF.

The model was developed on 49,314 women from the HFEA registry who started their second complete cycle between 1999 and 2008 using their own eggs and partner's sperm. As well as female age, number of eggs retrieved in the first complete cycle and the outcome of the first complete cycle (live birth, pregnancy loss, no pregnancy) were proven as key predictors (see Table 1). Other predictors included duration of infertility, tubal infertility, type of treatment and time between first and second egg retrievals. The model was externally validated on 39,442 UK women who underwent their second complete cycle between 2010 and 2016. The C-statistic was 0.65 and calibration showed a systematic overprediction of live birth for all women. The parameter estimates were recalibrated and subsequently the model showed much improved calibration. It should be noted that the validation data is now 6 years old, which may affect the accuracy of the model for new patients. Also, as mentioned earlier, the HFEA registry does not have some potentially important predictors.

According to the UK's National Institute for Health and Care Excellence guidelines, women under 40 years of age should be offered three complete cycles of IVF through the National Health Service [4]. However, since the local Clinical Commissioning Groups in the UK make their own decisions regarding access to IVF funding, this means that some parts of the country are offered anything from one to three fully funded complete cycles. Some are not provided any funding. Therefore, for many couples who do not have access to funding after one complete cycle this model will be particularly helpful as it can provide their predicted chance of live birth if they were to continue treatment. This will help them to plan ahead and prepare financially.

Important predictors of live birth

Knowing which characteristics are the most relevant for predicting live birth after IVF treatment is helpful for researchers wishing to either develop a model or, preferably, update existing models with new predictors that improve the performance. A systematic review of predictive factors in IVF by van Loendersloot et al [11] found that female age, duration of infertility, basal follicle stimulating hormone and number of oocytes were most relevant. However, the study called for better quality studies to focus on whether embryo quality and number of embryos transferred would be useful predictors.

McLernon et al [15] investigated the relative importance of each predictor in the two UK models. This was done by calculating the adequacy which is the proportion of the final model's goodness of fit (measured using the -2*log likelihood (-2LL) statistic) that is explained by the individual predictor [22-23]. For the final model (with all predictors included) the -2LL was calculated. Then the same statistic was calculated again for a model which is only adjusted for the complete cycle number and the particular predictor of interest (e.g., female age). The smaller model's -2LL is calculated as a proportion of the final model's -2LL. This is repeated for each of the remaining predictors. The predictor with the largest proportion is said to explain the most variation in the outcome. For the pre-treatment model, female age explained 85% of the total variation explained by all predictors. However, when treatment predictors were included, they found that female age (44%), cryopreservation of embryos (39%), and number of eggs (38%) each explained a similar high amount of the total variation explained by all the predictors. None of the other published IVF prediction models investigated adequacy or ranked predictors by importance.

However, a limitation of the adequacy method is that the proportion can appear large even if a predictor is weakly associated with the outcome. This would occur if the full model had a small -2LL i.e. doesn't explain much of the total variation. Furthermore, it will suffer from omitted variable bias which refers to important unknown or unavailable predictors of live birth which could potentially change the adequacy of another predictor. Steyerberg recommends that we simply judge the importance of each predictor by looking at the relative risk (i.e. odds

ratios) of the predictors and using clinical judgement [24]. In that respect, female age still comes out as the most important with an odds ratio (95% CI) of 1.66 (1.62 to 1.71) for a 37-year-old versus a 31 year old. Note, that age was not categorised, but was included in the model as restricted cubic spline terms to account for the non-linear relationship between age and the outcome. The odds ratio presented is the 25th percentile versus the 75th percentile value for age which is an easier way of interpreting the association for a non-linear relationship.

The SART pre-treatment model showed that female age, BMI and AMH had the strongest associations with live birth as did age, BMI and number of eggs collected for the post-treatment model [20]. BMI and AMH were unavailable in the UK database and so could not be included as predictors, however, duration of infertility was not available in the USA database. In all models, causes of infertility had reasonably small associations with live birth, with male factor, diminished ovarian reserve, uterine factor and tubal infertility having the strongest association.

When predicting from the second complete cycle, it is clear from the Ratna et al [21] study that number of eggs collected from the first retrieval and the outcome of the first complete cycle are also important to consider. For the latter predictor, the odds of a live birth for women who had a previous IVF live birth were almost twice that for women who had no pregnancy at all over the first complete cycle. Women who had a pregnancy loss (and no live birth) in the first complete cycle had a 35% increased odds of a live birth compared to women who had no pregnancy over the first complete cycle.

A note on useful complete cycle specific live birth prediction models

Two further models are worthy of note. Although they do not predict live birth cumulatively over multiple complete cycles of IVF, they do predict over the first complete cycle of IVF. While the following studies do not specifically use the term ‘complete cycle’ in their articles,

their approaches agree with our definition i.e. all fresh and subsequent frozen-thawed embryo transfer cycles from one episode of ovarian stimulation.

The Netherlands

The model by van Loendersloot et al [25] (identified by the Ratna et al (2020) [13] review) predicts the chance of ongoing pregnancy over the first complete cycle. It also predicts ongoing pregnancy at each successive complete cycle for couples in whom all previous complete cycles were unsuccessful. It was developed using a cohort of 1326 couples treated at a single centre in The Netherlands. The model was adjusted for the number of previous failed cycles as well as female age, duration of infertility, basal FSH, previous ongoing pregnancy and causes of infertility. It also includes predictors based on laboratory data from the previous failed IVF cycle e.g. fertilization method (IVF/ICSI), number of embryos after egg retrieval, mean morphological score per Day 3 embryo, presence of 8-cell embryos on Day 3 and presence of morulae on Day 3. It was externally validated on a dataset from the same centre but from a more recent time period. The c-statistic was 0.68 and the model was updated following evidence of miscalibration. Two further independent validation studies using data from single centres in Italy and Belgium showed lower discrimination (both 0.64) and poor calibration [26,27]. However, the Italian study recalibrated the model to find better agreement while the Belgian study fitted a new model to their own data.

We find that the van Loendersloot model is informative since it takes account of frozen-thawed cycles giving patients a fuller picture of their likely chances of success over their current complete cycle of treatment. We recommend further large external validation studies for this model since it was developed and validated on data from one centre. For use in other centres, external validation on data from those centres is recommended [28,29].

China

A model predicting cumulative live birth over the first complete cycle only was developed using data on almost 18,000 women from a University hospital in China [30]. Age, number of

oocytes, number of good quality embryos (defined as an embryo with 6–12 blastomeres graded 1 and 2), fertilization rate, treatment type (IVF versus ICSI) and duration of infertility were included as predictors. Age and oocytes were included as linear terms meaning that they did not adjust for the known non-linear relationship between these predictors and live birth [15,20]. The model was internally validated using 10 times 10-fold cross-validation, which resulted in a c-statistic of 0.74. The model has yet to be externally validated, but the final model parameters including the intercept were not presented which will make it difficult for independent investigators to conduct external validation on their datasets.

Recommendations and further work

For prediction of cumulative live birth over multiple complete cycles of IVF (where a complete cycle is defined as all fresh and frozen embryo transfers arising from a single episode of ovarian stimulation), we recommend the use of the UK and USA models at pre- and post-treatment. All were developed on national level datasets and followed the recommended reporting guidance for model development [19]. The pre-treatment models from both countries may be used before couples commence their first IVF cycle while the post-treatment models are useful before starting a second complete cycle [15,20,21]. However, it is not guaranteed that using these models in countries outside the UK and the USA will provide accurate predictions for their patients. Therefore all of these models need to be validated on independent geographical datasets for use in other clinics and countries. Further, they need to be continually validated and updated using data collected within the countries and clinics they have been developed in to prevent calibration drift [31]. Calibration drift can be caused by changes in casemix and IVF practice. Clinics or countries which display over or under prediction upon calibration assessment can still use the model after it has been recalibrated. This can be as simple as adjusting the model intercept to reflect the IVF success rates in that clinic or country. However, if that does not work there are several other ways of correcting miscalibration [32].

With respect to predictors that should be considered when developing new models or updating existing models, female age is the most important. Other factors that should be considered for prediction before starting treatment include duration of infertility, female BMI and markers of ovarian reserve. Ovarian reserve markers make a statistically and clinically significant contribution in the prediction of live birth following IVF treatment. However, they don't appear to be as important as female age (with which they correlate quite highly). Previous research seems to suggest that out of the ovarian reserve markers, AMH is the most reliable [33-35], and it has been shown to have some association with live birth independently of age [36]. However, another systematic review concluded that AMH and AFC added nothing when included with age in the prediction of ongoing pregnancy after IVF [18]. Future IVF prediction studies should utilise large (possibly national level) datasets with which to externally validate existing recommended models. They should investigate the added value of including different ovarian reserve markers to these models to confirm whether AMH is the most predictive.

For models that predict from the point of treatment, the number of eggs collected, double versus single fresh embryo transfer and blastocyst versus cleavage stage transfer should be considered. Further research is needed into whether embryo quality measures add further predictive accuracy. If sample size is not an issue, then it is important to include all known predictors, including those that are not strongly associated with live birth, to increase predictive accuracy. These include causes of infertility, previous pregnancy status and treatment type e.g. IVF versus ICSI.

Summary

IVF prediction models that estimate the chance of cumulative live birth over multiple complete cycle of treatment are useful to provide a complete picture of a couple's likelihood of success. Models have been developed using national level data in the UK and USA for

predictions before starting treatment. The UK has a further two models which provide revised predictions at later stages: one for use at the time of the first fresh embryo transfer and the other for use before starting a second complete cycle of treatment. The USA has one further model for use at the start of the second compete cycle but only for couples who first complete cycle was unsuccessful. Models developed using data from single centres in China and The Netherlands are able to predict pregnancy outcome over the first complete cycle. The latter can also be used to predict at each successive complete cycle assuming previous complete cycles failed. We recommend using the UK and USA models, but both need continual validation using updated patient data in order to be relevant in terms of predictive accuracy in new patients. All of the models require external validation in different geographical regions to ensure that they provide accurate predictions in those countries (or centres). For researchers developing new prediction models, the most important patient predictors to include are female age, duration of infertility, BMI and ovarian reserve markers. When revising predictions using treatment data, the model should include number of eggs collected. Further work is needed to determine the added predictive value of embryo quality.

Acknowledgments

None

Conflict of Interest

Conflicts of interest: none

Practice Points

- IVF prediction models should only be used at the intended moment of application, e.g. before IVF starts, and for the purpose in which they were intended to be used which is primarily for patient counselling and planning.
- Several models predicting cumulative live birth over complete cycles have been developed in the UK and USA, each of which is intended for a particular point in the patient's treatment e.g. before the first complete cycle starts, at the first embryo transfer, or before the start of the second complete cycle.
- Female age, duration of infertility, female BMI, and AMH are important pre-treatment predictors, while number of eggs collected adds value when treatment level information is known. More research is needed to assess others such as embryo quality and other ovarian reserve markers.

Research Agenda

- The UK and USA IVF prediction models should be validated using data from different geographical locations so that they can be used to make accurate predictions in patients undergoing treatment in those countries.
- Further studies are needed to find new or understudied predictors of IVF success such as embryo quality, markers of ovarian reserve and paternal age.
- Further studies are needed to determine the extent to which female age is related to markers of ovarian reserve.

References

1. De Mouzon J, Goossens V, Bhattacharya S, Castilla JA, Ferraretti AP, Korsak V, et al. Assisted reproductive technology in Europe, 2006: results generated from European registers by ESHRE. Hum Reprod 2010; 25: 1851–1862.

2. Roque M, Lattes K, Serra S, Solà I, Geber S, Carreras R, et al. Fresh embryo transfer versus frozen embryo transfer in in vitro fertilization cycles: a systematic review and meta-analysis. Fertil Steril 2013; 99: 156–162.

3. Cutting R. Single embryo transfer for all. Best Practice & Research Clinical Obstetrics & Gynaecology. 2018; 53: 30-37.

4. National Collaborating Centre for Women's and Children's Health (UK). Fertility: assessment and treatment for people with fertility problems. London: Royal College of Obstetricians & Gynaecologists; 2013 Feb. (NICE Clinical Guidelines, No. 156.) https://www.ncbi.nlm.nih.gov/books/NBK247932/.

5. Chambers GM, Dyer S, Zegers-Hochschild F, de Mouzon J, Ishihara O, Banker M, et al. International Committee for Monitoring Assisted Reproductive Technologies world report: assisted reproductive technology, 2014†. Hum Reprod 2021;36:2921-2934. doi: 10.1093/humrep/deab198. PMID: 34601605.

6. Luke B, Brown MB, Wantman E, et al. Cumulative birth rates with linked assisted reproductive technology cycles. N Engl J Med 2012;366:2483–91.

7. Macaldowie A, Wang YA, Chambers GM, Sullivan EA. National Perinatal Epidemiology and Statistics Unit, the University of New South Wales, Sydney. Assisted reproduction technology in Australia and New Zealand 2011. 2013. https://npesu.unsw.edu.au/surveillance/assisted–reproductive–technology–australia–new–zealand–2011.

8. Malizia BA, Hacker MR, Penzias AS. Cumulative live–birth rates after in vitro fertilization. N Engl J Med 2009;360:236–43.

9. McLernon DJ, Maheshwari A, Lee AJ, Bhattacharya S. Cumulative live birth rates after one or more complete cycles of IVF: a population–based study of linked cycle data from 178,898 women. Hum Reprod 2016;31:572–81.

10. Maheshwari A, McLernon D, Bhattacharya S. Cumulative live birth rate: time for a consensus? Hum Reprod 2015;30:2703–7.*

11. Van Loendersloot LL, van Wely M, Limpens J, Bossuyt PMM, Repping S, van der Veen F. Predictive factors in in vitro fertilisation (IVF): a systematic review and meta–analysis. Hum Reprod Update 2010;16:557–89.*

12. Steyerberg E. Clinical Prediction Models: a practical approach to development, validation, and updating. 2nd ed. New York, NY: Springer; 2019.

13. Ratna M, Bhattacharya S, Abdulrahim B, McLernon DJ. A systematic review of the quality of clinical prediction models in in vitro fertilisation. Hum Reprod 2020. doi: 10.1093/humrep/dez258.*

14. Luke B, Brown MB, Wantman E, Stern JE, Baker VL, Widra E, et al. A prediction model for live birth and multiple births within the first three cycles of assisted reproductive technology. Fertil Steril 2014;102:744–752.

15. McLernon DJ, te Velde E, Steyerberg E, Lee AJ, Bhattacharya S. Predicting the chances of a live birth after one or more complete cycles of in-vitro fertilisation: a population-based study of linked cycle data from 113,873 women. BMJ 2016;355:i5735. http://dx.doi.org/10.1136/bmj.i5735.*

16. Leijdekkers JA, Eijkemans MJC, van Tilborg TC, Oudshoorn SC, McLernon DJ, Bhattacharya S, et al, on behalf of the OPTIMIST group. Predicting the cumulative chance of live birth over multiple complete cycles of in vitro fertilisation: an external validation study. Hum Reprod 2018; doi:10.1093/humrep/dey263.

17. Van Calster B, Nieboer D, Vergouwe Y, de Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: From utopia to empirical data. Journal of Clinical Epidemiology. 2016;74:167–76. doi: 10.1016/j.jclinepi.2015.12.005.4

18. Broer SL, van Disseldorp J, Broeze KA, Dolleman M, Opmeer BC, Bossuyt P, 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.

19. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Annals of Internal Medicine. 2015;162(1):55–63. doi: 10.7326/M14-0697. PMID: 25560714.*

20. McLernon DJ, Raja EA, Toner JP, Baker VL, Doody KJ, Seifer DB, et al. Predicting personalized cumulative live birth following in invitro fertilization. Fertil Steril 2021;117:326-338. https://doi.org/10.1016/j.fertnstert.2021.09.015.*

21. Ratna M, Bhattacharya S, van Geloven N, McLernon DJ. Predicting cumulative live birth for couples beginning their second complete cycle of in vitro fertilisation treatment. Hum Reprod 2022;37:2705-86. https://doi.org/10.1093/humrep/deac152.11.*

22. Thompson D. Ranking Predictors in logistic regression, SAS Institute: Paper D10–2009. 2009. http://www.mwsug.org/proceedings/2009/stats/MWSUG-2009-D10.pdf.

23. Harrell FE. Regression modeling strategies with applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd Ed. Springer, 2015.

24. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925-31.

25. van Loendersloot LL, van Wely M, Repping S, Bossuyt PMM, van der Veen F. Individualized decision-making in IVF: calculating the chances of pregnancy. Hum Reprod 2013;28:2972–2980.*

26. Sarais V, Reschini M, Busnelli A, Biancardi R, Paffoni A, Somigliana E. Predicting the success of IVF: external validation of the van Loendersloot's model. Hum Reprod 2016;31:1245-52.

27. Devroe J, Peeraer K, Verbeke G, Spiessens C, Vriens J, Dancet E. Predicting the chance on live birth per cycle at each step of the IVF journey: external validation and update of the van Loendersloot multivariable prognostic model. BMJ Open 2020;10:e037289. doi:10.1136/bmjopen-2020-037289.

28. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. Journal of Clinical Epidemiology. 2016;69:245–7. doi: 10.1016/J.JCLINEPI.2015.04.005.

29. Altman DG, Royston P. What do we mean by validating a prognostic model? Statistics in Medicine. 2000;19:453–73. https://doi.org/10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5.*

30. Zhu H, Zhao C, Xiao P and Zhang S. Predicting the Likelihood of Live Birth in Assisted Reproductive Technology According to the Number of Oocytes Retrieved and Female Age Using a Generalized Additive Model: A Retrospective Cohort Analysis of 17,948 Cycles. Front Endocrinol 2021;12:606231. doi:10.3389/fendo.2021.606231.

31. Jenkins DA, Sperrin M, Martin GP, Peek N. Dynamic models to predict health outcomes: current status and methodological challenges. Diag Prognost Res 2018;2:23. https://doi.org/10.1186/s41512-018-0045-2.

32. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol 2008;61:76-86. doi: 10.1016/j.jclinepi.2007.04.018.

33. Tal R, Seifer DB. Ovarian reserve testing: a user's guide. Am J Obstet Gynecol 2017;217:129-140.

34. Toner JP, Seifer DB. Why we may abandon basal follicle-stimulating hormone testing: a sea change in determining ovarian reserve using antimullerian hormone. Fertil Steril 2013;99:1825-1830.

35. Barad DH, Weghofer A, Gleicher N. Comparing anti-Müllerian hormone (AMH) and follicle-stimulating hormone (FSH) as predictors of ovarian function. Fertil Steril 2009;91:1553-1555.

36. Iliodromiti S, Kelsey TW, Wu O, Anderson RA, Nelson SM. The predictive accuracy of anti-Müllerian hormone for live birth after assisted conception: a systematic review and meta-analysis of the literature. Hum Reprod Update 2014;20:560-70. doi: 10.1093/humupd/dmu003.

MCQs and answers with full explanations

Question 1

IVF prediction models that estimate the chance of a live birth over multiple complete cycles of treatment are clinically useful because:

a) They tell the clinician whether or not to offer the couple IVF treatment
b) They inform the couple and clinician their personalised chance of having a baby over the first complete cycle, and cumulatively over subsequent complete cycles
c) They provide estimates of treatment effect which help decide whether to have IVF or ICSI.
d) They help the couple manage their expectations and prepare emotionally and financially for IVF.
e) They can be used to decide when IVF treatment should begin.

Answers to question 1

(a) F (b) T (c) F (d) T (e) F

Explanation to answers to question 1

(a) These models only provide an estimate of the chance of live birth for couples starting IVF treatment. These chances are presented as a percentage risk. The model does not classify the predictions into dichotomised yes or no decisions. Furthermore, these models have been developed for use in couples who are about to undergo IVF. For models that are to be used to decide on whether or not a patient should have treatment, the dataset used for model development must also contain patients who never have treatment, and such patients must be considered in the model. This is not the case for the models presented here.

(b) The models have been adjusted for a couple's personal and treatment-based characteristics which provide a more individualised prediction of success. The model provides the predicted chance over the first complete cycle of treatment, where a complete cycle is defined as all fresh and frozen embryo transfers resulting from a single episode of ovarian stimulation. It is also provides the chance of live birth cumulatively over the first and second complete cycle, and so on.

(c) While these models have been adjusted for treatment type i.e. IVF versus ICSI, they are not intended to be used to decide between these two strategies. The reason is because retrospective data has been used to develop these models rather than randomised controlled trial data. The latter would concern a two-armed trial comparing live birth outcomes in patients following randomisation to ether IVF or ICSI. This is not the aim of these prediction models presented here and population-based data are prone to indication bias which concerns the bias caused by not appropriately adjusting for the reason why IVF or ICSI was undertaken in the patients used for model development e.g. preference of clinician or couple (details not available in the data used for modelling). Estimating treatment effectiveness using population-based data is not an easy thing to do and requires careful causal modelling.

(d) The model provides the predicted chance over the first complete cycle of treatment, and cumulatively over the first and second complete cycle, and so on. Such predictions are useful for couples to get a view of their likely chance of success in order to manage their

expectations. Knowing their predicted chance over multiple complete cycles will help them prepare financially for the number of complete cycles they may decide to undergo.

(e) These models have all been developed for use just before a defined point of treatment e.g. before ovarian stimulation or at first embryo transfer etc. The time to start of a particular stage of treatment is not incorporated. Prognostic models which provide the background chances of pregnancy or live birth without treatment may be useful to decide when treatment should be undertaken e.g. once the prediction goes below 30%.

Question 2

IVF prediction models should be used to calculate the cumulative chance of live birth over multiple complete cycles:

a) only at the intended moment of application.
b) only before the start of the first complete cycle since that is when couples enter the risk set.
c) at the start of any complete cycle because cumulative models include subsequent cycles.
d) only and not for prediction over a single complete cycle of treatment.
e) so that the clinician can decide whether or not the couple should have IVF treatment.

Answers to question 2

(a) T (b) F (c) F (d) T (e) F

Explanation to answers to question 2

(a) only at the intended moment of application because at later time points patient predictor values and the live birth prevalence will have changed. Also, further important predictors will be known from previous treatments. These changes will not be considered in the model which was developed using patient information at the original time point.

(b) That is when couples start their IVF treatment, but not necessarily when they enter the risk set. It depends when the particular model is designed to be applied to patients. For example, the model by Ratna et al, 2022 was developed to be used before the second complete cycle.

(c) It is true that cumulative models include subsequent cycles. However, models that predict the cumulative chance of live birth over multiple complete cycles must only be used to make predictions at the intended moment of use. For example, the McLernon et al, 2016 pre-treatment model predicts cumulative live birth over six complete cycles but only for patients about to begin their first complete cycle. However, there are models that can predict pregnancy outcomes over each successive complete cycle such as the model by van Loendersloot et al, but not cumulatively over multiple complete cycles.

(d) Models that predict the cumulative chance of live birth over multiple complete cycles can be used to predict live birth over the first complete cycle, and cumulatively over the first and second, then cumulatively over the first, second and third, etc. As stated in the explanation for part (c), there are models that can predict pregnancy outcomes over each successive complete cycle such as the model by van Loendersloot et al, but not cumulatively over multiple complete cycles.

(e) These models have been developed for use in couples who are about to undergo IVF. For models that are to be used to decide on whether or not a patient should have treatment, the dataset used for model development must also contain patients who never have treatment, and such patients must be considered in the model. This is not the case for the models presented here.

Table 1 Clinical prediction models predicting cumulative live birth for couples undergoing IVF including validation results and predictors

Time of use in patients	Study	Country of development	Outcome	Internal validation performance	External validation performance	Predictors
Before first ovarian stimulation	McLernon et al, 2016 [15]	UK	CLB up to six complete cycles	C=0.69 (95% CI 0.68 to 0.69) Calibration slope[1] = 0.996	Prospective cohort from The Netherlands (Leijdekkers et al. 2018) [16]: C=0.62 (95% CI 0.59 to 0.64) Calibration-in-the-large=-0.23 (95% CI -0.36 to -0.10); calibration slope=0.98 (95% CI 0.69 to 1.27) (after recalibration, the calibration plot showed improved accuracy of predictions) After updating model with AMH, AFC and body weight,	Complete cycle number (1 to 6), woman's age, duration of infertility, treatment type (ICSI versus IVF), Year first complete cycle started (for predictions in new patients this was always set to the latest year, 2009), tubal infertility, male factor infertility, unexplained infertility, anovulatory infertility, previous pregnancy in couple.

					C=0.66 (95% CI 0.64 to 0.68)[2]	
	McLernon et al, 2021 [20]	USA	CLB up to three complete cycles	C=0.71 (increasing to 0.73 when AMH included). Calibration plots show good agreement.	Not done	Complete cycle number (1 to 3), woman's age, previous full-term birth, male factor infertility, polycystic ovary syndrome, uterine factor, diminished ovarian reserve, unexplained infertility, woman's BMI, AMH (in secondary model only)
	Van Loendersloot et al, 2013 [25]	The Netherlands	Predicts ongoing pregnancy over the first complete cycle. Also, predicts for each successive	C=0.68 (95% CI 0.65 to 0.70). Calibration using Hosmer-Lemeshow test, p=0.41	Temporal validation on 440 couples by development team. C=0.68 (95% CI 0.63 to 0.73); Calibration slope=0.85 (95% CI 0.53 to 1.17) indicating slightly optimistic predictions. Model was recalibrated. External validation using data	Woman's age, duration of infertility, previous ongoing pregnancy, male factor infertility, diminished ovarian reserve, endometriosis, basal FSH, number of previous failed cycles, age X male infertility, endometriosis X diminished

			complete cycle assuming the previous complete cycles failed.		from 840 women from a clinic in Italy (Sarais et al, 2016) C=0.64 (95% CI 0.61 to 0.67); calibration slope=1.88 (95% CI 1.51 to 2.33) reflecting poor agreement, which improved after recalibration [26]. External validation in 591 couples in single Belgian clinic (Devroe at al, 2020). C=0.64 (95% CI 0.61 to 0.68); calibration slope=0.643 (95% CI 0.471 to 0.815). Model was refitted to validation data and not recalibrated [27].	ovarian reserve. Embryo parameters based on previous failed cycles: Embryo yes v no after ovum retrieval, number of embryos after ovum retrieval, mean morphological score all embryos day 3, 8-cell embryo yes v no on day 3, morulae yes v no on day 3.
After embryo development	Zhu et al, 2021 [30]	China	CLB over the first complete	C=0.7394 (10x10-fold cross-	Not done	Woman's age, number of oocytes, number of good quality

in first cycle			cycle	validation)		embryos (defined as 6–12 blastomeres graded 1 and 2), fertilisation rate, treatment type (IVF versus ICSI), duration of infertility
At first embryo transfer attempt	McLernon et al, 2016 [15]	UK	CLB up to six complete cycles	C=0.76 (95% CI 0.75 to 0.77); Calibration slope[1]=0.998	Prospective cohort from The Netherlands (Leijdekkers et al, 2018) [16]: C=0.71 (95% CI 0.69 to 0.74) Calibration-in-the-large=-0.01 (95% CI -0.12 to 0.11); calibration slope=0.97 (95% CI 0.77 to 1.19). After updating model with AMH, AFC and body weight, C=0.71 (95% CI 0.69 to 0.73)[2]	Complete cycle number (1 to 6), woman's age, duration of infertility, treatment type (ICSI versus IVF), Year first complete cycle started (for predictions in new patients this was always set to the latest year, 2009), tubal infertility, previous pregnancy in couple, cryopreservation of embryos in first complete cycle (yes v no), number of eggs collected in first complete cycle,

						stage of embryos transferred (no embryo transfer, single cleavage, single blastocyst, double blastocyst, triple cleavage, triple blastocyst versus double cleavage).
Before second ovarian stimulation	Ratna et al, 2022 [21]	UK	CLB from the second up to the fourth complete cycle	C=0.66 (95% CI 0.65 to 0.67)	Temporal validation on UK data (2010-2016) C=0.65 (95% CI 0.64 to 0.65); Calibration-in-the-large = -0.08; Calibration slope=0.85 (95% CI 0.81 to 0.88). Model was recalibrated by subtracting 0.22 and multiplying all regression coefficients by 0.85.	Complete cycle number (2 to 4), woman's age, duration of infertility, treatment type (ICSI versus IVF), Year second complete cycle started (for predictions in new patients this was always set to the latest year), tubal infertility, time between first and second egg retrieval (months), number of eggs retrieved at the first

						complete cycle, outcome of first complete cycle (live birth, pregnancy loss versus no pregnancy)
Before second ovarian stimulation in those whose first complete cycle did not result in a live birth	McLernon et al, 2021 [20]	USA	CLB from the second up to the third complete cycle	C=0.71; Calibration plots show good agreement	Not done	Complete cycle number (2-3), woman's age, male factor infertility, polycystic ovary syndrome, uterine factor, diminished ovarian reserve, woman's BMI, number of eggs collected at first complete cycle.

CLB=cumulative live birth

[1]Optimism-adjusted calibration slope calculated using non-parametric bootstrap (see supplementary text of McLernon et al. 2016) [15].

[2]Optimism adjusted c-statistic using non-parametric bootstrap (see Leijdekkers et al, 2018) [16].