Human Reproduction, Vol.24, No.8 pp. 1774–1778, 2009
Advanced Access publication on April 23, 2009 doi:10.1093/humrep/dep109

human reproduction
OPINION

Evaluating prediction models in reproductive medicine

S.F.P.J. Coppus[1,2,3,4], F. van der Veen[1], B.C. Opmeer[2], B.W.J. Mol[1,3], and P.M.M. Bossuyt[2]

[1]Department of Obstetrics and Gynaecology, Centre for Reproductive Medicine, Academic Medical Centre, Amsterdam, The Netherlands [2]Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, Room J1B-216-1, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands [3]Department of Obstetrics and Gynaecology, Máxima Medical Centre, Veldhoven, The Netherlands

[4]Correspondence address. E-mail: s.f.coppus@amc.uva.nl; s.coppus@mmc.nl

Prediction models are used in reproductive medicine to calculate the probability of pregnancy without treatment, as well as the probability of pregnancy after ovulation induction, intrauterine insemination or *in vitro* fertilization. The performance of such prediction models is often evaluated with a receiver operating characteristic (ROC) curve. The area under the ROC curve, also known as *c*-statistic, is then used as a measure of model performance. The value of this *c*-statistic is low for most prediction models in reproductive medicine. Here, we demonstrate that low values of the *c*-statistic are to be expected in these prediction models, but we also show that this does not imply that these models are of limited use in clinical practice. The calibration of the model (the correspondence between model-based probabilities and observed pregnancy rates) as well as the availability of a clinically useful distribution of probabilities and the ability to correctly identify the appropriate form of management are more meaningful concepts for model evaluation.

Key words: prediction model / fertility / spontaneous pregnancy / IUI / IVF

Introduction

In the evaluation of subfertile couples, clinicians have traditionally focused on finding the underlying cause of subfertility. Yet, only in a minority of couples, a causal diagnosis can be made, one which fully explains why conception has not occurred. Examples of such causal diagnoses are anovulation, azoospermia and bilateral tubal pathology. In the majority of couples, a relative factor is found, only partially explaining why conception has not occurred. The latter factors include advanced maternal age, mild male subfertility, cervical hostility, mild endometriosis and one-sided tubal pathology. In other couples, no factors impairing fertility can be found at all, and a diagnosis of unexplained subfertility is made.

A number of treatments options are available once a causal diagnosis has been made. In anovulatory women, infertility can be corrected by ovulation induction. Women with bilateral tubal disease can be treated with tubal surgery or offered *in vitro* fertilization (IVF). In men with azoospermia or oligoasthenozoospermia, the inability to conceive can be overcome by surgical sperm retrieval and assisted fertilization. Treatment of these couples is indicated, as their probability of natural conception is virtually zero. In contrast to couples with a causal diagnosis, couples with a relative factor as well as those with unexplained subfertility vary in their prognosis. Whereas some have relatively good chances of achieving a pregnancy without treatment, these chances are rather low in others (Collins *et al.*, 1995). As targeted treatments are not available for these conditions, empiric interventions such as intrauterine insemination (IUI) and IVF are usually offered. IUI and IVF are expensive and not without potential side effects, and these treatment options should therefore be offered to couples only if their probability of conceiving with treatment is sufficiently higher than their probability of conceiving without treatment.

Clinical decision making in these couples should therefore be guided by these probabilities (Habbema *et al.*, 2004). A number of validated prediction models are available to help the clinician in calculating the probability of pregnancy in subfertile couples, with and without treatment (see Leushuis *et al.*, 2009 for a critical appraisal). There are models to calculate: the probability of a treatment-independent pregnancy (Eimers *et al.*, 1994; Collins *et al.*, 1995; Snick *et al.*, 1997; Hunault *et al.*, 2004; Hunault *et al.*, 2005; van der Steeg *et al.*, 2007), the probability of a pregnancy after IUI (Steures *et al.*, 2004; Custers *et al.*, 2007) and the probability of a pregnancy after IVF (Stolwijk *et al.*, 1996; Templeton *et al.*, 1996; Smeenk *et al.*, 2000; Lintsen *et al.*, 2007).

For most of these models, the ability to predict pregnancy has been evaluated by receiver operating characteristic (ROC) curves. These curves are obtained by comparing the proportion of couples exceeding a pre-set probability threshold within those who achieve a pregnancy with the proportion also exceeding that threshold in those who do not achieve a pregnancy in a specified time frame.

This is done for all possible probability thresholds. The area under the resulting ROC curve (AUC), also known as *c*-statistic, expresses the extent to which a model can identify couples who will become pregnant and is used as a measure of model performance.

In general, the *c*-statistic for prediction models in reproductive medicine is rather low, ranging between 0.59 and 0.64 for models for treatment-independent pregnancy, between 0.56 and 0.59 for IUI and between 0.50 and 0.67 for IVF. A graph of such a typical ROC curve is shown in Fig. 1, adapted from a paper on the prediction of pregnancy after IUI (Custers *et al*., 2007). According to criteria for evaluating areas under ROC curves, these values indicate low-to-modest discriminatory performance (Swets, 1988). It is questionable whether the *c*-statistic is the best way to express the predictive performance of a model and, consequently, if a limited discriminatory capacity precludes the use of the model in clinical practice. In this paper, we illustrate that the area under the ROC curve or *c*-statistic has limited value in the evaluation of prediction models in reproductive medicine. We argue that calibration and prognostic classification are more relevant concepts for such models to guide clinical decision making.

c-Statistic in diagnostic research

In principle, the aim of diagnostic testing is to distinguish diseased from non-diseased patients. The diagnostic accuracy of a test is studied by comparing the result of the test under evaluation with the result of a reference standard in a series of patients, the latter being the best available method for classifying a patient as diseased or not. The results of the cross-classification can then be expressed as the sensitivity and specificity of the test under evaluation. The sensitivity, or true-positive fraction, reflects the proportion of diseased patients with a positive test result, whereas the specificity, or one minus the false-positive fraction, reflects the proportion of patients without the disease with a negative test result.

In case the studied test is of a continuous nature, the sensitivity and specificity of the test depend on the cut-off value to define positive and negative test results. Sensitivities and specificities can be calculated over the whole range of possible cut-off values, where higher sensitivities are obtained at lower specificities and visa versa. The resulting series of sensitivity–specificity pairs is plotted as an ROC curve, showing the sensitivity versus the specificity for each possible cut-off value. AUC indicates how well the test discriminates between diseased and non-diseased patients. An AUC of 1 implies perfect discrimination, whereas an AUC of 0.5 means that the test does not discriminate at all (Hanley and McNeil, 1982). Assuming an error-free reference standard, every diagnostic test has the potential of being 100% discriminative.

c-Statistic in prognostic research

Prediction models in reproductive medicine evaluate whether a combination of factors can predict the occurrence of pregnancy within a specified time period. In contrast to the situation in diagnostic testing, in which the disease is present at the moment of testing, pregnancy has not yet occurred at the time the potentially predictive factors are measured. Whether or not pregnancy occurs can be considered the outcome of a stochastic process occurring over time. This implies that the value of a test cannot be expressed by the proportion of patients who are pregnant at the time of testing, as this is essentially zero. The association has to be expressed in terms of the probability that pregnancy will occur in the future. This probability can be calculated for a single test or for a multivariable prediction model, the latter most often developed using Cox proportional hazard modelling. Such prediction models are often evaluated similar to diagnostic tests, using ROC curves and their AUC or *c*-statistic. In doing so, the calculated probability of pregnancy within a specified time period is regarded as the continuous test result, and the reference standard is the actual occurrence of pregnancy within that specified period of follow-up. Using different cut-off values for these probabilities yields a series of sensitivity and specificity pairs, through which an ROC curve can be drawn. For prediction models in reproductive medicine, the AUC or *c*-statistic reflects how well the model is able to distinguish those with a future pregnancy from those that will not achieve a pregnancy within the time frame specified, not the degree to which couples with a higher probability will conceive first.

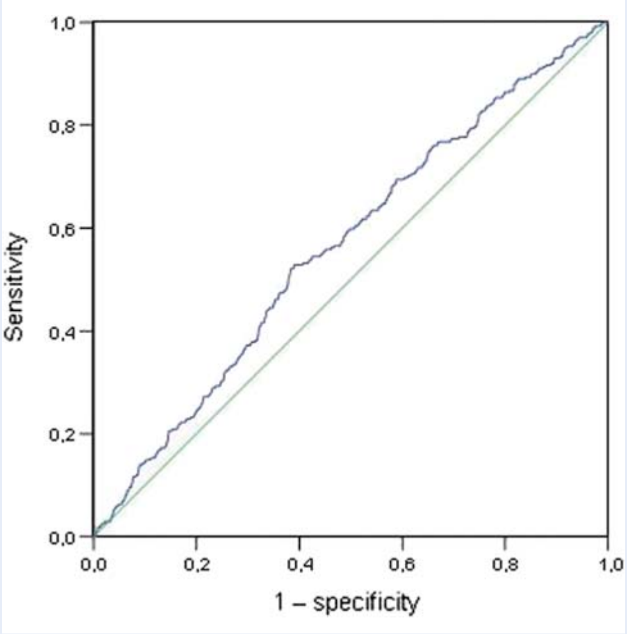

Figure 1 Typical ROC curve of a prediction model in reproductive medicine. This ROC curve derived from a paper in which a prediction model for pregnancy after IUI was externally validated, demonstrating an AUC of 0.56 (Custers *et al*., 2007). Sensitivity is defined here as the percentage of cycles not resulting in an ongoing pregnancy that was predicted correctly, and specificity was defined as the percentage of cycles that resulted in an ongoing pregnancy that was predicted correctly. Reprinted from Custers *et al*. (2007). Copyright 2007, with permission from Elsevier.

Although this may seem quite logical at prima facie, the crux of the matter is that perfect discrimination is achievable only if the population of subfertile couples consists of two distinct subpopulations: fertile couples that have not yet conceived but will do so in the near future and infertile couples guaranteed not to conceive. It is very unlikely that this is a fair representation of couples after an initial subfertility workup. Subfertility is not dichotomous in nature but a complex multifactorial phenomenon, corresponding to a gradual continuum

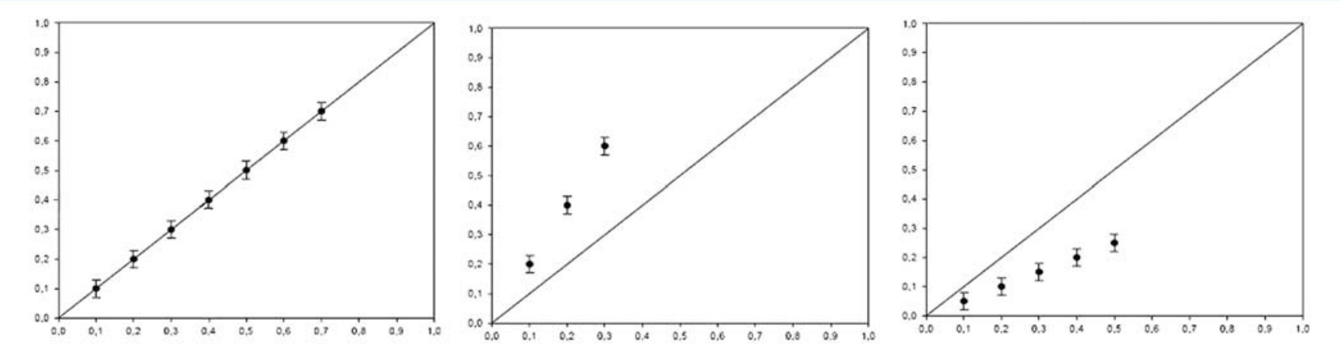

Figure 2 Calibration plots with calculated probability on the *X*-axis and observed proportion on the *Y*-axis. The left plot shows perfect calibration. The middle plot demonstrates a model that tends to give underestimated probabilities, whereas the plot on the right shows systematic overestimation.

of impaired reproductive capacity. Couples with very good pregnancy chances will rarely enter a subfertility workup, as they are the ones that usually have conceived spontaneously within 12 months after the start of unprotected intercourse. Most couples with a causal diagnosis, the ones with poorest fertility prospects, are identified in the first part of the subfertility workup and treated accordingly. With the two extremes eliminated, the remaining group of subfertile couples predominantly has an intermediate prognosis.

When perfect discrimination is not achievable, the maximum value of the *c*-statistic depends on the underlying distribution of probabilities. The lower the variability of the probabilities, the lower the maximum value of the *c*-statistic will be (Gail and Pfeiffer 2005; Cook 2007). Among subfertile couples, defined as couples that have not conceived despite a year of unprotected intercourse, the probability to conceive spontaneously within the next 12 months is more or less normally distributed in couples with unexplained subfertility, with a mean probability of 30% (van der Steeg *et al*., 2007). As the distribution of probabilities in couples that successfully conceive has considerable overlap with that of couples that are not successful, the maximum *c*-statistic that can be reached with any model is in the order of 0.62 (Gail and Pfeiffer, 2005; Cook, 2007).

Calibration

So if pregnancy itself cannot be predicted with 100% accuracy, what then? We should realize that the modern approach in subfertility treatment is to offer tailored management to individual couples, in which treatment is offered only to couples who have poor chances of a spontaneous pregnancy. Ideally, early treatment in couples with high chances of natural conception should be avoided, as should the delay of treatment in couples with poor prospects of pregnancy. Such an approach could reduce costs and may prevent multiple pregnancies as well as other complications of assisted reproductive technologies (Steures *et al*., 2006). For this tailored treatment, a classification of couples is required. This classification will not be in terms of those that will become pregnant versus those that will not, as this is almost impossible. The relevant classification is one in terms of those for whom the probability of becoming pregnant with treatment, be it IUI or IVF, is sufficiently higher than the probability of becoming pregnant without treatment, versus those for which this is not the case.

Prediction models in reproductive medicine should be judged on this capability. One necessary condition is that we should be able to trust calculated probabilities. The reliability of these model-based probabilities can be evaluated by inspecting the calibration of the model: the degree to which calculated probabilities agree with actual observed event rates. The level of calibration of a prediction model can be explored by assigning couples to subgroups, based on the calculated probabilities, and then, after follow-up, comparing the average calculated probability in each subgroup with the proportion of couples that became pregnant. One way of assigning couples to subgroups is using deciles of calculated probabilities. A comparison of the mean probability in each subgroup versus the observed proportion can be plotted in a calibration plot, as in Fig. 2. In case of perfect calibration, all points in the calibration plot will be located on the diagonal, the line of equality. In the left plot, probabilities are well calibrated, whereas the middle and right plots in Fig. 2 show calculated probabilities that suffer from underestimation and overestimation, respectively.

A combination of perfect calibration and perfect discrimination can only be achieved with perfect prediction: a probability of 100% for all those with a future event, and a probability of 0% for all others. Such an extreme bimodal distribution is not very realistic in most applications, including subfertility (Ware 2006; Cook 2007). In fact, there is a trade-off between calibration and discrimination. A simulation model showed that, assuming a realistic range of probabilities, the *c*-statistic for a perfectly calibrated model would be 0.83 at best (Diamond, 1992). Overlapping probability distributions in those with and those without the future event will usually prevent that maximum from being reached (Gail and Pfeiffer, 2005).

Prognostic classification

Although good calibration is very important, a well-calibrated model is not necessarily useful in subfertility care. Probabilities also need to have sufficient variability over a clinically useful range, as the essence of prediction models is their ability to correctly classify individuals into clinically useful risk strata (Cook *et al*., 2006; Cook, 2007; Ridker and Cook, 2007). Imagine a perfectly calibrated model that assigns a probability of achieving a treatment-independent pregnancy of 20–30% to all subfertile couples, and a second perfectly calibrated model that assigns a probability of IVF success of 20–30% in the same couples. That would probably lead to a recommendation of expectant management in all, which is likely to be erroneous, and the models would not allow us to identify couples that are better off with IVF, separating them from couples that are better off with expectant management. For models to be useful, they need to not only be well calibrated, but also produce a clinically useful distribution in probabilities in the target population. If we believe that some couples are better off with IVF, both the spread of probabilities with expectant management and the spread of probabilities with IVF should be sufficiently wide and distributions should be well apart. Only then couples can be subdivided into prognostic categories to decide on the appropriate treatment policy. Various graphs and expressions of the range in calculated probabilities have recently been suggested in the literature. One example is the predictiveness curve, suggested by Pepe *et al*. (2007), which plots the cumulative distribution of the calculated probabilities in a group, and as such is a useful visual aid in expressing the relative variability in calculated probabilities.

Research is emerging, and a debate has started, on the best ways to evaluate superiority of models for prediction (Pepe *et al*., 2007; Janes *et al*., 2008; Pencina *et al*., 2008). In most of these discussions, a change in the *c*-statistic is not regarded as the best statistic to measure an improvement in model performance (Cook, 2007), as a substantial improvement in the *c*-statistic can only be achieved by unrealistically large associations between the predictor variables and outcome (Ware, 2006). Whereas odds ratios for diagnostic tests exceeding 30 or more are of no exception, odds ratios for prognostic tests rarely exceed 2 (Glas *et al*., 2003). Most importantly, a comparison of areas under the curve between prognostic models does not show us whether individual couples have a different prognosis in one model when compared with a second model. Regarding the latter, some have suggested examining how many couples are reclassified, i.e. reassigned from the group of those for whom expectant management is the better option to those for whom, say, IUI or even IVF is the better alternative. Ultimately, however, the value of prediction models should be based on outcome: to what extent does the use of models improve patient outcome, i.e. allow a couple to have a healthy pregnancy, at a reasonable balance with patient burden, morbidity and costs.

Conclusion

In reproductive medicine, prognostic models that perfectly predict pregnancy in subfertile couples, after natural conception or after assisted reproductive technologies, do not exist and most likely will never exist. Yet, properly calibrated models, with sufficient variability in calculated probabilities, can be used to support clinical decision making when the chances of a pregnancy with treatment have to be weighed against the chances of a pregnancy without treatment. The commonly used *c*-statistic or area under the ROC curve expresses only discrimination and is, as such, not a good measure of the extent to which predictive markers and models can guide decision making. To assess the clinical value of prediction models, calibration, variability in probabilities and, subsequently, the degree to which reclassification of these probabilities yield clinically relevant prognostic strata are more relevant criteria.

Funding

This work was supported by grant 91746364 in the VIDI program The Netherlands Organisation for Health Research and Development (ZonMW) and by grant 92003519 in the MD-medical research trainee program.

References

Collins JA, Burrows EA, Wilan AR. The prognosis for live birth among untreated infertile couples. *Fertil Steril* 1995;**64**:22–28.

Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;**115**:928–935.

Cook NR, Buring JE, Ridker PM. The effect of including c-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med* 2006;**145**:21–29.

Custers IM, Steures P, van der Steeg JW, van Dessel TJ, Bernardus RE, Bourdrez P, Koks CA, Riedijk WJ, Burggraaff JM, van der Veen F *et al.* External validation of a prediction model for an ongoing pregnancy after intrauterine insemination. *Fertil Steril* 2007;**88**:425–431.

Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. *J Clin Epidemiol* 1992;**45**:85–89.

Eimers JM, te Velde ER, Gerritse R, Vogelzang ET, Looman CW, Habbema JD. The prediction of the chance to conceive in subfertile couples. *Fertil Steril* 1994;**61**:44–52.

Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. *Biostatistics* 2005;**6**:227–239.

Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003; **56**:1129–1135.

Habbema JD, Collins J, Leridon H, Evers JL, Lunenfeld B, te Velde ER. Towards less confusing terminology in reproductive medicine: a proposal. *Hum Reprod* 2004;**19**:1497–1501.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.

Hunault CC, Habbema JD, Eijkemans MJ, Collins JA, Evers JL, te Velde ER. Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models. *Hum Reprod* 2004;**19**:2019–2026.

Hunault CC, Laven JS, van Rooij IA, Eijkemans MJ, te Velde ER, Habbema JD. Prospective validation of two models predicting pregnancy leading to live birth among untreated subfertile couples. *Hum Reprod* 2005;**20**:1636–1641.

Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med* 2008;**149**:75–760.

Leushuis E, van der Steeg JW, Steures P, Bossuyt PM, Eijkemans MJ, van der Veen F, Mol BW, Hompes PG. Prediction models in reproductive medicine: a critical appraisal. *Hum Reprod Update* 2009; in press.

Lintsen AM, Eijkemans MJ, Hunault CC, Bouwmans CA, Hakkaart L, Habbema JD, Braat DD. Predicting ongoing pregnancy chances after IVF and ICSI: a national prospective study. *Hum Reprod* 2007;**22**:2455–2462.

Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;**27**:157–172.

Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. *Am J Epidemiol* 2007;**167**:362–368.

Ridker PM, Cook NR. Biomarkers for prediction of cardiovascular events. *N Engl J Med* 2007;**356**:1474–1475.

Smeenk JM, Stolwijk AM, Kremer JA, Braat DD. External validation of the Templeton model for predicting success after IVF. *Hum Reprod* 2000; **15**:1065–1068.

Snick HK, Snick TS, Evers JL, Collins JA. The spontaneous pregnancy prognosis in untreated subfertile couples: the Walcheren primary care study. *Hum Reprod* 1997;**12**:1582–1588.

Steures P, van der Steeg JW, Mol BWJ, Eijkemans MJ, van der Veen F, Habbema JD, Hompes PG, Bossuyt PM, Verhoeve HR, van Kasteren YM *et al*. Prediction of an ongoing pregnancy after intrauterine insemination. *Fertil Steril* 2004;**82**:45–51.

Steures P, van der Steeg JW, Hompes PG, Habbema JD, Eijkemans MJ, Broekmans FJ, Verhoeve HR, Bossuyt PM, van der Veen F, Mol BW. Intrauterine insemination with controlled ovarian hyperstimulation versus expectant management for couples with unexplained subfertility and an intermediate prognosis: a randomised clinical trial. *Lancet* 2006;**368**:216–221.

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.

Swets J. Measuring the accuracy of diagnostic systems. *Science* 1988; **240**:1285–1293.

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

Van der Steeg JW, Steures P, Eijkemans MJ, Habbema JD, Hompes PG, Broekmans FJ, van Dessel HJ, Bossuyt PM, van der Veen F, Mol BWJ. Pregnancy is predictable: a large-scale prospective external validation of the prediction of spontaneous pregnancy in subfertile. *Hum Reprod* 2007;**22**:536–342.

Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med* 2006;**355**:2615–2617.