Human Reproduction, Vol.31, No.9 pp. 1931–1933, 2016
Advanced Access publication on July 12, 2016 doi:10.1093/humrep/dew170

human reproduction

INVITED COMMENTARY

Clinical decision-making in azoospermic men: in search of the ideal prediction model

David J. McLernon and Siladitya Bhattacharya*

Institute of Applied Health Sciences, University of Aberdeen, Foresterhill, Aberdeen AB25 2ZD, UK

*Correspondence address. E-mail: s.bhattacharya@abdn.ac.uk

Submitted on June 2, 2016; resubmitted on June 2, 2016; accepted on June 9, 2016

Clinical prediction models estimate the individual chance of an existing condition (diagnostic) or of a future health event (prognostic) in patients, using information about their personal characteristics, history, test results and/or treatment. In this issue of *Human Reproduction*, two separate papers report on the development of prediction models in severe male infertility. The first aims to predict successful testicular sperm extraction (TESE) in men with non-obstructive azoospermia (Cissen *et al.*, 2016), while the second estimates the chances of live birth in couples undergoing ICSI using surgically extracted sperm (Meijerink *et al.*, 2016).

What do these models mean for fertility clinicians and their patients? To answer this question we need to consider not only the validity of these models and the methods used to generate them but also the clinical context in which these models might be used.

Recent guidelines have been introduced to help improve the way prediction models are reported (Collins *et al.*, 2015). The papers published in this issue display considerable insight in terms of planning and execution of the analysis. In the model predicting live birth following ICSI, where the number of events is modest ($n = 224$), the authors have had to deal with the challenge posed by the relatively large array of candidate predictors by using a forward selection procedure with generous entry criteria (e.g. *P*-value <0.2) and exit criteria (e.g. *P*-value >0.2). Such an approach has been shown to increase power for the selection of variables that appear to have a weak effect in a small dataset but are actually well known to be predictive (Steyerberg *et al.*, 2000). In the second model on sperm recovery after TESE, the authors investigated the relationship between continuous variables such as serum LH, FSH and testosterone levels and successful sperm recovery using a model that assumes a linear association between these predictors and the outcome. It is worth noting that serum LH levels showed a non-linear relationship and this was appropriately accounted for using a polynomial transformation (Steyerberg, 2009). A positive aspect of both models is that each has been validated on a separate dataset from another centre—an essential step before they can be implemented in clinical practice (Bleeker *et al.*, 2003).

The presence of missing data in several of the predictor variables has been acknowledged in both papers. In the model predicting sperm recovery with TESE, the authors used multiple imputation based on an assumption that any systematic difference between the missing values and observed values can be explained by differences in the observed characteristics of the patients (Sterne *et al.*, 2009; van Buuren, 2012). This is better than only using patients with complete data as such an approach leads to a loss of power and introduces bias, especially if there are systematic differences between the complete and incomplete cases. Whilst it is good that the authors have considered missing data they provide limited information about the imputation modelling method used and the plausibility of the missing at random assumption as recommended in recent guidelines (Sterne *et al.*, 2009). For example, were there any important differences between patients with complete and missing data? How many imputed datasets were created? Which predictors were included in the imputation process?

The performance of a prediction model can be evaluated by a variety of methods. Two of the most important are discrimination (i.e. does the model assign a higher predicted probability to those who have the outcome versus those who do not?) and calibration (how well do predicted probabilities compare with observed probabilities?). The former can be assessed using the area under the receiver operating curve whilst the latter can be assessed using calibration plots (Steyerberg, 2009). Both have been described—but to what extent do they tell us whether these prediction models will actually improve decisions in clinical practice (Steyerberg *et al.*, 2010)?

Both decision models address various stages of treatment of azoospermia—TESE in the first paper and ICSI using surgically retrieved sperm in the second. The models developed have area under the curve (AUC) values of 0.62–0.69, i.e. they offer a slightly better discriminatory capacity than pure chance (AUC = 0.5) but are far from offering 100% sensitivity or specificity in a situation where the alternative to surgery/ICSI is the use of donor sperm or no pregnancy at all. Implicit in both papers is the belief that providing an estimate of the chances of success may help patients make an informed choice. For many men, the

consequences of the use of their own gametes versus donor gametes are so profound that the threshold for attempted surgical sperm retrieval is likely to be quite low. The calibration plot for the sperm extraction model showed that it could reliably distinguish between men with a 35 and 50% chance of retrieval and those whose chances were <35%. Many would argue that a figure <35% might still be acceptable to a lot of patients, given that similar rates are quoted to many couples attempting assisted reproduction. TESE itself is a relatively minor operation usually performed under local anaesthetic. Thus it is likely that some men might choose surgery despite there being a poor chance of success. Live birth rates associated with IVF have always been modest and are by no means the only outcome of relevance to patients. Women undergoing assisted reproduction, many of whom may never have a child, have tended to value all (health and non-health) attributes of a particular treatment rather than simply a live birth outcome (Ryan, 1999). Thus, false positives and false negatives could be valued very differently by azoospermic men and their partners when decisions are made around TESE. Additionally, once sperm have been recovered, it is unusual to see a couple who decide to avoid an autologous ICSI attempt in favour of treatment with donor sperm or no treatment at all.

One way of navigating this complex decisional dilemma is by using a decision curve analysis approach. For example, the hypothetical example in Fig. 1 allows us to determine a range of threshold predicted probabilities of obtaining sperm with TESE by plotting the net benefit against the threshold probability. The net benefit is defined as the difference between the proportion of men who are true positive and false positive for successful sperm retrieval, weighted by the relative harm of a false-positive and false-negative result (Vickers and Elkin, 2006; Vickers *et al.*, 2016). The net benefit of the model is then compared with the scenario where everyone is assumed to have no chance of obtaining sperm (the black dot and dash line in Fig. 1) and therefore the procedure is not done (net benefit equals zero) and to the scenario where everyone is assumed to have obtainable sperm and therefore undergo the procedure (the blue line in Fig. 1). Threshold probabilities that have a higher net benefit than both of these scenarios (the red dashed in Fig. 1) means that, for these thresholds, use of the model is better than either withholding or offering TESE to the whole population. To determine whether the model is of clinical value, we need to consider the possible range of threshold probabilities of obtaining sperm at which the clinician would decide to perform the TESE, taking into account the background 50% success (Tournaye, 2010), the minimally invasive nature of the operation and the alternative, i.e. donor sperm. This might result in a probability threshold of, say, 10–20% at which surgery could be offered.

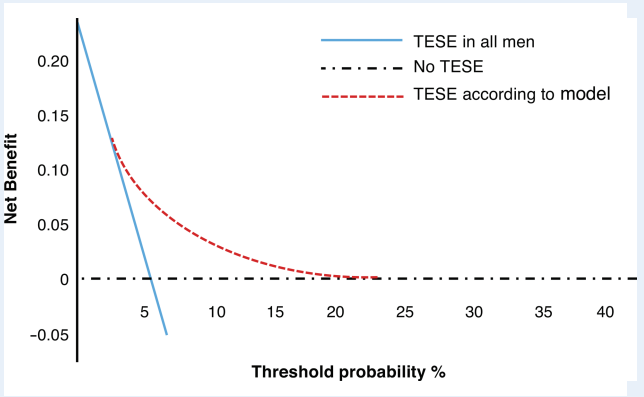

Figure 1 A hypothetical decision curve of net benefit of testicular sperm extraction (TESE) in men with non-obstructive azoospermia.

Prediction models cannot replace the role of clinicians in terms of making a decision regarding treatment or prognosis. They provide additional information which when placed alongside other medical facts facilitates clinical decision-making (Moons *et al.*, 2009). However, very rarely are any prediction models implemented into clinical practice. The impact of a model on clinical pathways, health outcomes and costs need to be investigated and shown to be beneficial before doctors and their patients can be persuaded to use them (Reilly and Evans, 2006; Moons *et al.*, 2012).

This is particularly true for reproductive medicine. A systematic review of prediction models in this field found that out of 29 models, of which 8 were externally validated (three showed good performance), only one was ever used in clinical practice (Leushuis *et al.*, 2009). This is an area which needs attention. Cluster randomized trials, where a clinic is randomly assigned to standard care only or the additional use of a prediction model, need to be considered (Moons *et al.*, 2012).

To be useful in men with azoospermia, a prediction model may need to ensure either that its false negative rate is close to zero or that it is able to dichotomize clearly between those who have any sperm versus those who have none. Patients currently agree to TESE knowing it can fail in half of all cases, and accept ICSI which offers an even lower chance of live birth per cycle. For prediction models to be widely accepted in this population, we need robust evidence to show that their use will lead to more benefit than harm.

Authors' roles

Both authors contributed towards developing the concept, drafting the manuscript and approving the final version.

Funding

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

Conflict of interest

None declared.

References

Bleeker SE, Moll HA, Steyerberg EW, Donders ART, Derksen-Lubsen G, Grobbee DE, Moons KGM. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol* 2003;**56**:826–832.

Cissen M, Meijerink AM, D'Hauwers KW, Meissner A, van der Wiede N, Mochtar MH, de Melker AA, Ramos L, Repping S, Braat DDM *et al.* Prediction model for obtaining spermatozoa with TESE in men with non-obstructive azoospermia. *Hum Reprod* 2016; doi:10.1093/humrep/dew147.

Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med* 2015;**13**:1.

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

Meijerink AM, Cissen M, Mochtar MH, Fleisher K, Thoonen I, de Melker AA, Meissner A, Repping S, Braat DDM, van Wely M *et al*. Prediction model for live birth in ICSI using testicular extracted sperm. *Hum Reprod* 2016; doi:10.1093/humrep/dew146.

Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *Br Med J* 2009;**338**:1317–1320.

Moons KGM, Pascal Kengne A, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;**98**:691–698.

Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Int Med* 2006;**144**:201–209.

Ryan M. Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilisation. *Soc Sci Med* 1999;**48**:535–546.

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

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

Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small datasets. *Stat Med* 2000;**19**:1059–1079.

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**:128–138.

Tournaye H. Update on surgical sperm recovery – the European view. *Hum Fertil* 2010;**13**:242–246.

van Buuren S. *Flexible Imputation of Missing Data*. Boca Raton: CRC Press, 2012.

Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;**26**:565–574.

Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models. *Br Med J* 2016;**352**:i6.