Human Reproduction, Vol.34, No.3 pp. 584–586, 2019
Advanced Access publication on January 11, 2019

human reproduction
LETTERS TO THE EDITOR

Accurate prediction of the irrelevant remains irrelevant

Sir,

We read with interest the recently published study in Human Reproduction—'Prognostic models reliably predict low and high ovarian response to stimulation' (Scheinhardt *et al.*, 2018). The study concluded that the outcome of ovarian stimulation with 150 μg corifollitropin alfa in a fixed, multiple dose GnRH-antagonist protocol can be validly predicted using logistic regression models with AMH being of paramount importance. While we congratulate the authors on their paper we are concerned about the usefulness of the study.

First, we question the use of low ovarian response (LOR) as outcome measure as the clinical relevance of this outcome is negligible. Even if one can accurately predict the patient's individual probability of being a low responder, no treatment has been shown to improve the prognosis for these women (Pandian *et al.*, 2010; Nagels *et al.*, 2015; Lensen *et al.*, 2018). We question if we should assess ovarian reserve with parameters such as AFC and AMH in all women starting IVF. Why bother the burden and costs of these tests if they do not positively influence the primary outcome of interest to the patient: the chance of having a baby? It is unfathomable that in 2018, more than a decade after a debate on this topic (Heijnen *et al.*, 2004; Min *et al.*, 2004), many published studies still report intermediate outcomes which lack clinical impact and are therefore irrelevant.

In contrast to the aforementioned, we do understand the authors choosing a high ovarian response (HOR) as their safety outcome, as it can help clinicians decide which women should not be prescribed corifollitropin alfa 150 ug. Still we wonder whether the presented prediction model provides sufficient benefit as the authors state that using the suggested 'HOR 1' model, corifollitropin alfa 150 ug treatment might not be considered appropriate if an individual's predicted HOR probability exceeds 28%. We think many clinicians and patients would feel this threshold is rather high, as in 2018 an OHSS free clinic should be one of our main aims. This could be achieved by lowering the FSH dose in women predicted to be at increased risk of HOR (Nyboe Andersen *et al.*, 2017; Oudshoorn *et al.*, 2017). Alternatively, potentially equally effective strategies are using a GnRH antagonist for downregulation with a GnRH agonist trigger if a HOR occurs and a subsequent freeze all of all embryos. This latter strategy is also possible in women treated with corifollitropin alfa 150 ug, despite the fact that when using this compound lowering the dose for ovarian stimulation is not possible.

We commend the authors' objective to externally validate the performance of the Oehringer models although it is unclear why they chose single test accuracy measures instead of the more informative *c*-statistic and calibration plot. It is also puzzling why they still go on to develop new predictive models which is particularly striking as these new models hardly outperform the Oehringer models (e.g. AIC differs with <2 points) and some are more difficult to use in daily practice (e.g. requiring log transformation of AMH). All in all, we feel it would have been better if the authors had chosen the conventional validation route (starting with the existing model, external validation using the c-statistic and calibration plot and then re-calibration if necessary) (Collins *et al.*, 2014), in particular as the selected 'HOR 1' and 'LOR1' models will now actually still require external validation in another dataset, before they can be considered robust enough to introduce into clinical practice.

Finally, we wonder why the authors chose to do a complete case analysis, which is known to introduce a risk of bias and reduce precision instead of performing multiple imputation (Janssen *et al.*, 2009). Importantly, in this dataset only 32 events of LOR and 25 of HOR were recorded. Using the general rule of thumb for model building (that states that you should consider one potential predictor for every 10 outcomes of interest) the authors may have seriously increased the risk of spurious findings by including nine potential predictive factors.

Summarizing, this article likely has methodological issues and reports the irrelevant outcome LOR. Predicting a HOR may have clinical impact, but only if FSH dose reduction is considered.

Conflict of interest

H.LT. received an unrestricted personal grant from Merck BV. F.J.M.B. receives monetary compensation as a member of the external advisory board for Merck Serono (the Netherlands) and Ferring pharmaceutics BV (the Netherlands), for advisory work for Gedeon Richter (Belgium) and Roche Diagnostics on automated AMH assay development, and for a research cooperation with Ansh Labs (USA). B.W.M. is supported by a NHMRC Practitioner Fellowship (GNT1082548) and reports consultancy for ObsEva, Merck and Guerbet.

References

Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu LM, Moons KG *et al*. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;**19**:14–40.

Heijnen EM, Macklon NS, Fauser BC. What is the most relevant standard of success in assisted reproduction? The next step to improving outcomes of IVF: consider the whole treatment. *Hum Reprod* 2004;**9**:1936–1938.

Janssen KJM, Vergouwe Y, Donders ART, Harrell FE, Chen Q, Grobbee DE, Moons KGM. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem* 2009;**55**:994–1001.

Lensen SF, Wilkinson J, Leijdekkers JA, La Marca A, Mol BWJ, Marjoribanks J, Torrance H, Broekmans FJ. Individualised gonadotropin dose selection using markers of ovarian reserve for women undergoing in vitro fertilisation plus intracytoplasmic sperm injection (IVF/ICSI). *Cochrane Database Syst Rev* 2018;**2**:CD012693.

Min JK, Breheny SA, MacLachlan V, Healy DL. What is the most relevant standard of success in assisted reproduction? The singleton, term gestation, live birth rate per cycle initiated: the BESST endpoint for assisted reproduction. *Hum Reprod* 2004;**19**:3–7.

Nagels HE, Rishworth JR, Siristatidis CS, Kroon B. Androgens (dehydroepiandrosterone or testosterone) for women undergoing assisted reproduction. *Cochrane Database Syst Rev* 2015;**11**:CD009749.

Nyboe Andersen A, Nelson SM, Fauser BCJM, García-Velasco JA, Klein BM, Arce J-C, ESTHER-1 study group. Individualized versus conventional ovarian stimulation for in vitro fertilization: a multicenter, randomized, controlled, assessor-blinded, phase 3 noninferiority trial. *Fertil Steril* 2017;**107**:387–396.

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

Pandian Z, McTavish AR, Aucott L, Hamilton MP, Bhattacharya S. Interventions for 'poor responders' to controlled ovarian hyper stimulation (COH) in in-vitro fertilisation (IVF). *Cochrane Database Syst Rev* 2010:CD004379.

Scheinhardt MO, Lerman T, König IR, Griesinger G. Performance of prognostic modelling of high and low ovarian response to ovarian stimulation for IVF. *Hum Reprod* 2018. doi:10.1093/humrep/dey236. [Epub ahead of print].

H.L. Torrance[1,*], F.J.M. Broekmans[1], and B.W.J. Mol[2]
[1] *Department of Reproductive Medicine and Gynaecology, University Medical Centre Utrecht, Utrecht University, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands*
[2] *Department of Obstetrics and Gynaecology, Monash University, Scenic Blvd & Wellington Road, Clayton VIC 3800, Australia*

*Correspondence address. Department of Reproductive Medicine and Gynaecology, University Medical Centre Utrecht, Utrecht University, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands.
E-mail: h.torrance@umcutrecht.nl

doi:10.1093/humrep/dey379
Advanced Access Publication on January 11, 2019

Reply: Ovarian response and its prediction are relevant

We thank our colleagues Dr Torrance, Dr Broekmans and Dr Mol for their interest in our study. In their letter, Torrance *et al.* question the value of predicting ovarian response in principle. This comes as a surprise, given that all three authors have themselves extensively been publishing on the prediction of high and low ovarian response (LOR) to ovarian stimulation (Hendriks *et al.*, 2005; Verhagen *et al.*, 2008; Broer *et al.*, 2013a, 2013b; Broekmans *et al.*, 2014; Hamdine *et al.*, 2015). While Torrance and Broekmans testify in their own work that 'accurate prediction of ovarian response prior to IVF is important' (Hamdine *et al.*, 2015) and that 'the predictability of ovarian response categories in antagonist co-treatment cycles is an important finding' (Broekmans *et al.*, 2014), they now have apparently come to the conclusion that ovarian response prediction is futile and irrelevant. This is even more surprising when considering that the same authors indeed utilize ovarian response prediction in order to design and conduct large interventional studies in both predicted low and predicted high responders (OPTIMIST trial: Oudshoorn *et al.*, 2017, van Tilborg *et al.*, 2017).

Regarding the clinical utility of predicting LOR we agree with Torrance *et al.* that at present no treatment has been shown to improve the prognosis for women with poor ovarian response. However, future interventional studies on new treatment options for LOR should use established prediction models for targeting subsets of patients by ovarian response. For daily clinical practice, prediction of LOR may furthermore be important in order to reduce treatment intensity and cost in those women, in which the therapeutic window for exogenous FSH stimulation has closed. Predicting a high ovarian response (HOR) to 150 μg corifollitropin alfa is undoubtedly of clinical relevance, since an FSH dose reduction can reduce burden, costs and risks without compromising outcomes (Nyboe Andersen *et al.*, 2017; Oudshoorn *et al.*, 2017). Regarding the cut-off, at which an individual should be considered at risk of HOR, we want to highlight that, firstly, only a small fraction of patients with HOR will indeed develop OHSS (Griesinger *et al.*, 2016), and secondly and more importantly, that our models allow the estimation of the HOR risk of an individual patient. This in turn allows the physician to make a case-by-case clinical judgment considering other patient specific risk factors. Finally, there is an obvious ex-ante interest of patients and doctors in the prospects of treatment by ovarian response (Sunkara *et al.*, 2011).

Regarding the methodological issues mentioned in their letter, we appreciate a fruitful discussion on the usage of well-suited methods. Some of their suggested methods definitely depict reasonable alternatives. Still, we consider some points as overly critical or unwarranted.

Torrance *et al.* raised the question why we 'chose single test accuracy measures instead of the more informative *c*-statistic and calibration plot'. We would like to point out that we explicitly presented area under the receiver operating characteristic curve (AUC) values—also called *C*-statistic or *C*-index—for all the compared models in Table II of our article.

They were puzzled why we 'still go on to develop new predictive models… [which] hardly outperform the Oehninger models'. While we openly discussed the non-superiority of the new models for the prediction of HOR, we did observe an improvement in the discriminative ability of the new models for the prediction of LOR. More importantly, novel models have to be developed in order to find out whether they can outperform the original ones, and the results of these scientific endeavors should be reported regardless of the outcome. On a side note, the main improvement of the LOR models was achieved by means of the log transformation of anti-muellerian hormone (AMH) levels, which is commented to be 'difficult to use in daily practice' by Torrance *et al.* We respectfully disagree on this point, computing a logarithm and an exponential function can be done with a simple hand-held calculator.

We wholeheartedly agree with Torrance *et al.* that the conventional validation route starts with an existing model, uses external validation and re-calibration if required, before entering clinical practice. We emphasize that in scientific practice, this is a continuing process that includes refinements and improvements of existing models. In that vein, our paper presents an external validation of the Oehninger models and at the same time suggests improvements of existing models that will, as we explicitly state, have to be externally validated in further studies.

They also raised the question why we 'chose to do a complete case analysis, which is known to introduce a risk of bias and reduce precision, instead of performing multiple imputation'. We agree that using multiple imputation of missing values might have been one way to go. However, complete case analysis might lead to biased results only if missing values are not missing completely at random, for which we did not see any indication. Certainly, complete case analysis comes at the