Added value of ovarian reserve testing on patient characteristics in the prediction of ovarian response and ongoing pregnancy: an individual patient data approach

Simone L. Broer1,2‡, Jeroen van Disseldorp1,2‡, Kimiko A. Broeze1,2, Madeleine Dolleman1,2, Brent C. Opmeer1,2, Patrick Bossuyt1,2, Marinus J.C. Eijkemans1,2, Ben-Willem J. Mol1,2, and Frank J.M. Broekmans1,2 on behalf of the IMPORT study group**

1Department of Reproductive Medicine, University Medical Center Utrecht, Room F05.126, P.O. Box 85500, Utrecht 3508 GA, The Netherlands 2Department of Obstetrics and Gynaecology, Amsterdam Medical Center, Amsterdam, the Netherlands

*Correspondence address. Tel: +31-88-755-1044; Fax: +31-88-755-5433; E-mail: S.L.Broer-2@umcutrecht.nl, www.ipd-meta-analysis.com/ipd

**A full list of authors is provided in the Appendix section.

Submitted on November 5, 2011; resubmitted on July 10, 2012; accepted on July 30, 2012

TABLE OF CONTENTS

- Introduction
- Methods
  - Data acquisition
  - Statistical analysis
- Results
  - Data acquisition
  - Prediction of a poor response or ongoing pregnancy from patient characteristics
  - Prediction of a poor response or ongoing pregnancy from ovarian reserve tests
  - Multivariable prediction models for poor response and ongoing pregnancy
  - Accounting for FSH dosage and study quality
- Discussion

BACKGROUND: Although ovarian reserve tests (ORTs) are frequently used prior to IVF treatment for outcome prediction, their added predictive value is unclear. We assessed the added value of ORTs to patient characteristics in the prediction of IVF outcome.

METHODS: An individual patient data (IPD) meta-analysis from published studies was performed. Studies on FSH, anti-Müllerian hormone (AMH) or antral follicle count (AFC) in women undergoing IVF were identified and authors were contacted. Using random intercept logistic regression models, we estimated the added predictive value of ORTs for poor response and ongoing pregnancy after IVF, relative to patient characteristics.

RESULTS: We were able to collect 28 study databases, comprising 5705 women undergoing IVF. The area under the receiver-operating characteristic curve (AUC) for female age in predicting poor response was 0.61. AFC and AMH each significantly improved the model fit (P-value < 0.001). Moreover, almost a similar accuracy was reached using AMH or AFC alone (AUC 0.78 and 0.76, respectively). Combining

† Both authors contributed equally.
Introduction

The incorporation of ovarian reserve tests (ORTs) in IVF management started after initial publications indicated a potential role for basal FSH in predicting pregnancy outcome after IVF and in counseling patients (Mahesh et al., 1988; Scott et al., 1989). Since these first publications, a large body of additional work on basal FSH and several other tests has been published, often with inconsistent findings on the magnitude and direction of the predictive effect. It became evident that the clinical value of previously published prediction models was highly dependent on the consequences related to the prediction (i.e. counseling versus refraining from treatment). Moreover, female age, which is strongly related to IVF outcome, was frequently omitted as a prime contributor in the prediction models (Broekmans et al., 2006; Verhagen et al., 2008).

Overall, individual studies have shown considerable variation in the predictive capacity of ORTs. The conventional way to summarize the available evidence would be to perform a systematic review and meta-analysis of the sensitivity and specificity of ORT, as reported in published studies (Leeflang et al., 2008). However, a major problem in interpreting these studies is the striking heterogeneity in individual patient populations, stimulation protocols, hormone assays ultrasound techniques and other features. Conventional meta-analysis of the accuracy of tests cannot easily account for this heterogeneity, nor does it respect the continuous nature of ORT data, or the statistical dependence between related tests and variables, i.e. ORT results are related to female age, and both are predictive of IVF outcome (Broeze et al., 2009).

To arrive at summary estimates of the added value of ORTs in women undergoing IVF, we undertook a meta-analysis with original individual patient data (IPD). By collecting test results, age and other patient characteristics, and IVF outcome in each individual patient, we would be able to respect the continuous nature of ORT data and would be able to study the added value of ORT to basic patient characteristics in predicting IVF outcome. Our aim was to answer the question of whether the most widely used ORTs, follicle stimulating hormone (FSH), antral follicle count (AFC) and anti-Müllerian hormone (AMH) added significantly and substantially to baseline female characteristics, such as age, in predicting the outcome of IVF treatment.

Methods

Data acquisition

We searched the literature for studies on the value of FSH, AFC and AMH in predicting IVF outcome. We built on searches performed in previous, conventional systematic reviews on the subject (Broekmans et al., 2006; Broer et al., 2009). A systematic search was performed in Medline to identify additional eligible papers, published until December 2009 (Fig. 1). Eligible for the current review were studies presenting data on at least one ORT and at least one patient characteristic and IVF outcome, in terms of ovarian response to stimulation, clinical or ongoing pregnancy or both.

Key words were used as synonyms for in vitro fertilization (IVF, controlled ovarian stimulation, in vitro fertilization) and synonyms for the respective ORT (FSH, Follicle Stimulating Hormone, AFC, Antral Follicle Count or number, AMH, Anti-Müllerian Hormone, Müllerian inhibiting substance). All titles and abstracts were evaluated for eligibility by two authors (S.B., J.D.) and if necessary the opinion of a third author was decisive (F.B.).

All authors of identified potentially eligible primary studies were informed about this IPD meta-analysis project and invited to share their data in a collaborative project. If authors were inclined to participate, they were provided with a data request form, informing them of the format of the data requested.

After data acquisition, all data were carefully examined and when possible converted into a single format. Any issues or inconsistencies were checked with the original author. For a more detailed description of our IPD meta-analysis methodology, the reader is referred to previous papers of our group (Broeze et al., 2009, 2010).

A comparison was made between the studies that could and those that could not be included. Sensitivity and specificity pairs at a certain threshold level for the prediction of a poor response or ongoing pregnancy were calculated for included and not included studies. A Spearman correlation was then calculated for sensitivity and specificity to specify that the differences in sensitivity and specificity levels between included and not included studies were only the result of the different threshold levels used.

We evaluated the quality of the included studies using the QUADAS checklist, supplemented by a number of items to evaluate the risk of bias in prognostic studies.

Statistical analysis

All analyses were performed for predicting both poor response as well as ongoing pregnancy after IVF as the outcome of interest. Poor response was defined as the yield of four or fewer oocytes at follicle aspiration or a cancelled cycle due to poor ovarian response (<3–4 dominant follicles >12 mm diameter), since this is a commonly used definition for poor response (Broer et al., 2009). Ongoing pregnancy was defined as a visible gestational sac on ultrasound with heartbeat at a gestational age of at least 9 weeks.

Duration of subfertility was defined as the period from the cessation of oral contraceptive use or start of unprotected intercourse until the first IVF attempt. Patients were stimulated according to the local protocol. In almost all studies, a starting dosage of at least 150 IU of FSH was used, which is the optimal daily dosage in expected normal responders. With this dose, it can be assumed that all patients receive maximum stimulation,
creating growth of all follicles sensitive to FSH in that time frame (Sterrenburg et al., 2011). We evaluated whether the ORT and patient characteristics, female age, BMI and duration of subfertility, were missing in the individual study databases. Whenever a particular variable was missing in an individual database, we made no attempt at imputation.

Random intercept logistic regression prediction models were then created with the ‘Lme4’ library in R version 2.9.0. (http://www.r-project.org/), using the Laplace approximation to the likelihood. These models were created to quantitatively estimate the added value of the ORTs on the patient characteristics in predicting poor response or ongoing pregnancy.

The random intercept model takes heterogeneity into account by assuming that included studies are a random sample of a potential universe of studies and that between-study variation in the predictive effect in this universe can be described by a normal distribution on the log odds scale. The model provides an estimate of the summary predictive effect as well as of the standard deviation of this distribution.

Three different sets of models for the prediction of poor response or ongoing pregnancy were used. The first model set included the patient characteristics female age, BMI and duration of subfertility. In the second set of models, the predictive capacities of the individual ORTs, FSH, AFC or AMH, in combination with predictive patient characteristics, were estimated. In the third set of models, the added value of combinations of ORTs to the patient characteristics was evaluated.

To account for between-study differences and their potential effect on our conclusions, we repeated the analyses using starting FSH dosage as a covariate. In similar analyses, we included study design features, as identified by the QUADAS checklist, as covariates in our models, to evaluate whether accounting for study design differences attenuated the observed associations between ORT, patient characteristics and the respective outcomes.

We constructed receiver-operating characteristic (ROC) curves to express the predictive accuracy of each of the combinations of predictive variables: the ability of the model to distinguish poor responders from the rest, and the ability to distinguish successful IVF couples from the rest. With each of the random intercept logistic regression models, we calculated the probability of the outcome of interest (poor response or ongoing pregnancy). By moving a positivity threshold from 0 to 1, we could then calculate sensitivity–specificity pairs for each model. Based on these, we plotted stratified ROC curves, with the ROC regression model as proposed by Janes et al. (2009) and Pepe et al. (2009). This model assumes that studies share a common ROC but allows the positivity threshold corresponding to each sensitivity–specificity pair to vary between studies. With this model the improvement in predictive accuracy of adding an ORT to other variables can be studied, while correcting for the heterogeneity between studies. This way we could compare the ROC and area under the curves (AUCs) of the models described above and evaluate them for statistically significant differences.

Because not all studies in this meta-analysis had reported data for all three ORTs, we constructed the prediction models using those databases from the total dataset that included the three ORTs (FSH, AFC and AMH) and age, to allow for direct comparisons, minimizing bias from indirect comparisons. The results of our analyses in the subgroup of three-test studies are shown in the main text, the result of the total study group are shown in the Supplementary Files.

Data were analyzed using SPSS 17.0 (SPSS Inc., Chicago, IL, USA), SAS 9.1 (SAS Institute Inc., Cary, NC, USA) and R version 2.9.0. (http://www.r-project.org/).
Results

Data acquisition

We identified 115 eligible study reports, from which we obtained contact information from 93 authors. Of these 93 authors, 67 replied to our (repeated) email or phone contact. Ultimately, we received 28 study databases that had been used for the publication of 55 manuscripts, provided by 24 collaborating authors (Tomas et al., 1997; Bancsi et al., 2000; Ng et al., 2000, 2005; Smeenk et al., 2000; van Rooij et al., 2002; Kwee et al., 2003; Popovic-Todorovic et al., 2003a, b; Yong et al., 2003; Erdem et al., 2004; Muttukrishna et al., 2004, 2005; Vladimirov et al., 2004, 2005; Ashrafi et al., 2005; van Swieten et al., 2005; Eldar-Geva et al., 2005; Klinkert et al., 2005a, b; Caroppi et al., 2006; Jayaprakasan et al., 2007; La Marca et al., 2007; Luna et al., 2007; Mclveen et al., 2007; Merce et al., 2007; Nelson et al., 2007; Smeenk et al., 2007; Liu and Greenblatt, 2008). These 28 databases contained data on 5705 subfertile women (Fig. 1). Data from 4170 women were suitable for poor response analysis, of these 893 (21%) had a poor response. Data from 5367 women could be used for the analysis of ongoing pregnancy prediction, of these 1231 women (23%) obtained an ongoing pregnancy.

Baseline characteristics of the 5705 women in the study group are summarized in Table I. The baseline characteristics of the original studies show some variation between the original studies, as do the poor response and pregnancy incidences, and the ORM averages (Supplementary data, Table SII). Study quality characteristics as scored by the QUADAS checklist and supplemental questions are shown in Fig. 2.

With the original data, we were able to replicate the primary findings of the original study in 10 databases. In 11 databases, the study database we received contained a number of patients that differed from the publication, whereas in 7 other databases there were slight inconsistencies in the baseline data previously published. The level of consistency between the individual data and the data reported in the published manuscript was considered sufficient for all included studies. No significant differences were found between the Spearman correlations of the included and not included studies for each ORM and outcome measure, indicating that the included and not included studies were comparable.

For 3235 women, both outcome measures were available and ongoing pregnancy rates for poor and normal responders could be compared. In normal responders, 30.2% of the women achieved an ongoing pregnancy compared with 11.7% of poor responders. Differences in ongoing pregnancy rates were also compared in age categories, demonstrating that poor responders have lower pregnancy rates than normal responders across all age groups, although this effect was gradually smaller in the higher age groups; over the age of 40 years comparable pregnancy prospects could be observed (Supplementary data, Table SII).

Prediction of a poor response or ongoing pregnancy from patient characteristics

For model building, we could use the data from 617 women for poor response analysis and from 420 women for ongoing pregnancy analysis. Of all patient characteristics, age was the strongest single predictor of poor response [odds ratio (OR) 1.12; 95% confidence interval (CI) 1.08–1.17; Supplementary data, Table SII]. BMI and duration of subfertility were not significantly predictive of poor response. In pregnancy prediction, age was the strongest single predictor of pregnancy, compared with other patient characteristics (OR 0.94; 95% CI 0.89–0.99) (Supplementary data, Table SII). Duration of subfertility was found not to be significantly associated with ongoing pregnancy, but BMI was. In a multivariable model, only BMI added any predictive value to age (Supplementary data, Table SII). Since age was the single constant and strongest predictor of poor response and ongoing pregnancy, we evaluated the added predictive effect of the ORM s FSH, AFC and/or AMH relative to the predictive value of age only in all further multivariable analyses.

Prediction of a poor response or ongoing pregnancy from ORMs

We compared the ORMs using the random intercept logistic regression model in predicting poor response (Table II). The ROC regression analysis demonstrated a high accuracy for AMH (AUC 0.78; 95% CI 0.72–0.84) and for AFC (AUC 0.76; 95% CI 0.70–0.82) in predicting poor response, but only a moderate accuracy for FSH (AUC 0.68; 95% CI 0.61–0.74; Table III). In predicting pregnancy after IVF, all three ORM s had only a very small or no predictive effect (Table II). The AUC were 0.53, 0.50 and 0.55 for FSH, AFC and AMH, respectively (Table III).

Multivariable prediction models for poor response and ongoing pregnancy

The multivariable analyses for poor response prediction showed that a model with age, AFC and AMH had a significantly higher predictive accuracy than a model based on age alone (AUC 0.80 versus 0.61; P < 0.001). Adding FSH to this model did not significantly improve predictive accuracy (P = 0.45; Table III). The predictive value of the multivariable model, including age and the two ORMs, AMH and AFC, was not significantly better than that of a single ORM, when used in isolation (P = 0.17 for AMH (AUC 0.78); P = 0.99 for AFC (AUC 0.76)). AMH, as a single predictor, has an accuracy comparable with all multivariable models with AMH and

| Table I Baseline characteristics of the 5705 women in the study group. |
|-------------------------------|-----------------------------|
| Patient characteristics       | Mean          (5th–95th percentile) |
| Female age (years)            | 34.3 (26.7–41.9) |
| FSH (IU/l)                    | 7.8 (3.8–14.0)  |
| AFC (number)                  | 11.6 (3.0–25.0) |
| AMH (ng/ml)                   | 2.1 (0.1–6.0)   |
| BMI (kg/m²)                   | 23.2 (18.5–30.1) |
| Duration of subfertility (years) | 4.01 (1.0–9.1) |
| Prevalences                   |                            |
| Poor response                 | 21.4%          |
| Ongoing pregnancy             | 22.9%          |

Poor response: ≤4 oocytes retrieved. Ongoing pregnancy: positive heartbeat at AD >9 weeks. Duration of subfertility: the period from the cessation of oral contraceptive use or start of unprotected intercourse until the first IVF attempt.
Age or with any of the other two ORT. The ROC curves for the multivari-
able models are shown in Fig. 3A.

Age was the strongest single predictor of pregnancy after IVF, with
moderate accuracy (AUC 0.57). Multivariable analysis for prediction of
ongoing pregnancy indicated that no single or combined ORT signifi-
cantly added predictive power to age (Table III). The AUC for the
combination of age, AMH and AFC was 0.59. ROC curves for the
multivariable analyses are shown in Fig. 3B.

Accounting for FSH dosage and study quality

In the prediction of a poor response, FSH dosage had a significant pre-
dictive effect (OR 1.009, P < 0.001). A higher FSH dosage was asso-
ciated with higher chances of a poor response. When FSH dosage was
added to the multivariable models of age and ORTs, the associations
of age and ORTs with poor response were very similar to those of the
models without accounting for FSH dosage. In the prediction of an
ongoing pregnancy, FSH dosage did not have a significant effect (OR
0.997, P 0.140). Here also, inclusion in the multivariable models did
not change the associations for age and ORTs. When included in
the multivariable models, none of the evaluated study quality charac-
teristics affected the predictive capacity of age and ORTs in predicting
poor response or ongoing pregnancy.

Discussion

The results of this IPD meta-analysis, based on 28 studies previously
reported, demonstrate that both AFC and AMH clearly add value
to female age in the prediction of poor ovarian response in IVF. Com-
parably good predictions can be made with either AMH or AFC alone,
without using female age. For the prediction of ongoing pregnancy
after IVF, ORTs do not add to the limited predictive capacity of
female age.

In the long-lasting debate on the true value of ovarian reserve
testing prior to IVF, a systematic review of the literature with
meta-analysis can be of help as an objective and systematic approach
in summarizing the available evidence. A major strength of the collabor-
ative effort reported here is its ability to analyze the independent
added value of several relevant predictors in a large body of data.

With the generous help of a large group of contributors, we have
been able to collect data on a number of patients, which far surpasses
that of the largest study performed so far, although it does not cover
the entire evidence base. Thereby, we have achieved consistency in
variable coding and a form of statistical analysis that accommodates
the remaining heterogeneity between studies.

Some potential limitations of our study have to be acknowledged.
For our analyses, the databases of only 55 of the eligible 115 manu-
scripts could be obtained. We were unable to reach a number of authors,
primarily because of inaccurate contact information, or
because authors did not reply on the e-mail addresses provided in
the study reports. Furthermore, older data were often lost, or kept
in a format that could no longer be read or could not be converted.
The Spearman correlations of the included and not included studies
were calculated and compared in order to study whether these
groups of studies were comparable. For none of the ORTs with
Table II Unvariable and multivariable models of age and ORT in the prediction of poor response and ongoing pregnancy.

<table>
<thead>
<tr>
<th></th>
<th>Poor response prediction</th>
<th>Ongoing pregnancy prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td>95% CI</td>
</tr>
<tr>
<td>Univariable models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (per year)</td>
<td>1.12</td>
<td>1.08–1.17</td>
</tr>
<tr>
<td>FSH (per IU/l)</td>
<td>1.27</td>
<td>1.19–1.35</td>
</tr>
<tr>
<td>AFC (per N)</td>
<td>0.77</td>
<td>0.73–0.82</td>
</tr>
<tr>
<td>AMH (per ng/ml)</td>
<td>0.50</td>
<td>0.41–0.60</td>
</tr>
<tr>
<td>Multivariable models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age and FSH</td>
<td>0.320</td>
<td></td>
</tr>
<tr>
<td>Age (per year)</td>
<td>1.12</td>
<td>1.07–1.17</td>
</tr>
<tr>
<td>FSH (per IU/l)</td>
<td>1.26</td>
<td>1.18–1.34</td>
</tr>
<tr>
<td>Age and AFC</td>
<td>0.192</td>
<td></td>
</tr>
<tr>
<td>Age (per year)</td>
<td>1.07</td>
<td>1.02–1.11</td>
</tr>
<tr>
<td>AFC (per N)</td>
<td>0.78</td>
<td>0.74–0.83</td>
</tr>
<tr>
<td>Age and AMH</td>
<td>0.321</td>
<td></td>
</tr>
<tr>
<td>Age (per year)</td>
<td>1.08</td>
<td>1.03–1.13</td>
</tr>
<tr>
<td>AMH (per ng/ml)</td>
<td>0.54</td>
<td>0.44–0.66</td>
</tr>
</tbody>
</table>

Results of random intercept logistic regression model in the prediction of poor response or ongoing pregnancy. For the prediction of a poor response, the multivariable analyses showed that all three ORT add predictive information to female age alone. Female age is the strongest predictor of ongoing pregnancy. All three ORT show a very small or absent predictive effect in the prediction of an ongoing pregnancy. Multivariable analyses show that all three ORT do not add predictive information to female age alone in the prediction of an ongoing pregnancy. P-values reflect whether the variable plays a significant role in the model. The column ‘Variance RI’ denotes the estimated variance of the random intercept in the random intercept logistic model. Its square root is the estimated standard deviation (SD), and may be interpreted on the logistic scale. A 1 SD difference in the population of studies corresponds to an increase in the Odds on the outcome (poor response and ongoing pregnancy, respectively) of exp (SD).

either outcome measure, a significant difference in the Spearman correlations was found. We therefore believe that the included and not included studies are comparable and that the current number of participants and level of detail allowed us to analyze a representative selection of the collected data.

The findings from our analyses confirm those of previous systematic reviews and meta-analysis of both single ORTs and multivariable prediction models for poor response to ovarian stimulation (Broekmans et al., 2006; Verhagen et al., 2008; Broer et al., 2009). Both AMH and AFC strongly represent the size of the cohort of FSH sensitive follicles continuously present in the ovaries, often referred to as the quantitative ovarian reserve. Response to ovarian hyperstimulation has been shown to be directly linked to this cohort size (Kwee et al., 2003). The role of AMH in marking the ovarian ageing process has been demonstrated in several studies showing that AMH decreases gradually with age and may be also predictive of the timing of menopause (de Vet et al., 2002; van Rooij et al., 2004, 2005; Sowers et al., 2008; van Disseldorp et al., 2008; Broer et al., 2011). From these data, the capacity of AMH as a marker for the quantitative ovarian reserve has become established.

For ongoing pregnancy prediction, age is the single most important predictor, although the accuracy in pregnancy prediction remains far from optimal. In contrast to their performance in predicting poor response, ORTs perform poorly in predicting pregnancy, as shown in this analysis. The large body of data in the present analyses finally clarifies the lack of added value of currently known ORTs to knowing female age. Since ovarian response to controlled hyperstimulation reflects quantitative ovarian reserve and the occurrence of an ongoing pregnancy after IVF is mainly related to qualitative ovarian reserve, it can be emphasized that ORTs reflect the quantitative aspect of the ovarian reserve status only. Qualitative ovarian reserve appears much harder to evaluate. In addition, ovarian reserve may not be the only factor affecting pregnancy chances in IVF/ICSI. Several factors, such as embryo quality, transfer technique or endometrial receptivity may be important (Boomsma and Macklon, 2006). It is likely that only by studying several consecutive treatment cycles, a true representation of a woman’s remaining reproductive capacity, based on her ovarian reserve status, may be obtained. Over the past decades, only one study evaluated the predictive role for ORTs in a series of subsequent IVF cycles, demonstrating that female age was the only factor predicting ongoing pregnancy after three treatment cycles, with no apparent role for ORTs (Hendriks et al., 2005).

The performance of assisted reproduction technology (ART) in infertile couples is far from optimal. Out of every 100 couples initiating IVF, only 50–60 will achieve their goal, even after having undergone several treatment cycles (Lintsen et al., 2007; Malizia et al., 2009). This high failure rate could be attributed to several factors, of which drop out rates and reduced ovarian reserve are the most popular ones. The urge to improve ART performance puts a high focus on identifying adequate ORTs. The limited accuracy of current tests has led to the situation that unfavorable test outcomes only lead to counselling and treatment adaptations that lack a solid scientific basis, instead of a refusal to offer ART treatment in the first place. Recent
studies have suggested a role for the use of patient characteristics, especially female age, combined with AMH for identification of various prognosis categories, with the poorest group having a per-cycle chance of pregnancy of 5% with a rather wide precision interval (1–16%; La Marca et al., 2011; Nelson and Lawlor, 2011). The question of how these predictions could alter patient management or aid in upgrading ART performance has remained unanswered. This may also be explained by the fact that very poor prognosis categories are difficult to identify with sufficient precision.

Recent publications have suggested the calculation of age-specific decline curves in order to maximize ORT accuracy (Barad et al., 2007; Henne et al., 2008; Scott et al., 2008). One study calculated age-specific FSH levels and live birth probabilities and demonstrated that variation in the chances of live birth is primarily determined by age and only to a lesser degree by basal FSH (Henne et al., 2008).

The analysis also demonstrated that FSH decline curves for five age groups yielded different cut-off values in the prediction of delivery rates (Scott et al., 2008). Since there was a very low rate of abnormal tests especially in the young age categories, the authors’ conclusion that age-specific basal FSH testing could serve as a reliable prognostic tool may be too optimistic.

It is to be expected that similar issues in the evaluation of tests and markers can be resolved with the meta-analysis of IPD. More and more funding agencies are inviting investigators to have a data sharing policy, and to allow others to benefit from the resources invested in the research. Inspired by the major successes achieved by the multicenter genetic consortia, those interested in clinical research could develop similar initiatives for patient centered research. We strongly believe that joining efforts in multicenter collaborations, possible even fine-tuning and coordinating study protocols through prospective meta-analysis, is an inevitable next step for clinical science in the 21st century, not just for randomized trials of interventions, but also in the evaluation of medical tests and biomarkers.

The clinical use of markers like AMH, basal FSH and the AFC is mostly based on cut-off levels. From the individual patient dataset,
across the studies, thereby prohibiting the calculation of relevant cut-off levels. To some extent, correction factors to standardize the results from various studies could be applied. Currently, however, this approach has not yielded final data for one of the three tests of interest. Therefore, centers for ART, applying tests for poor response prediction should rely on their own data analyses for cut-off level assignment. Indeed, development of centre-based prediction models for patient management or counseling is now gaining rapid attention (Banerjee et al., 2010).

Another possible issue may stem from the question of whether the poor responder patient observed in studies is indeed a genuine poor responder or merely the victim of insufficient FSH dosing. Response to ovarian hyperstimulation will mainly depend on the number of follicles sensitive to FSH in a certain time period. With dosages of at least 150 IU, the vast majority of patients will allow all such follicles to develop into dominant follicles. A dose–response effect seems to be only present in dosages under 150 IU daily (Sterrenburg et al., 2011). For the women with a small number of FSH sensitive follicles, applying high dosages, such as 300 or 450 IU would seem ineffective (Harrison et al., 2001; Klinkert et al., 2005a; Lekamge et al., 2008), although some studies offer some hope in this respect (Popovic-Todorovic et al., 2003a, b). Indeed, applying massive dosages of FSH may even affect the quality of the oocytes obtained from the responding follicles (Check et al., 2007). However, to evaluate the effect of FSH dosage in the prediction of a poor response and ongoing pregnancy, FSH dosage was added to the multivariable models. Although FSH dosage had a significant role in the models for the prediction of a poor response, it did not alter the predictive capacity of age and ORTs. Importantly a higher starting dosage of FSH was associated with a poor response. So, although FSH dosage affects the prediction of a poor response, the predictive effect may well be based on verification bias and therefore represent a spurious relation. Regarding the identified predictors, ORTs and female age, it can now be underlined that ovarian stimulation has been maximal for the great majority of cases included in this analysis. The occurrence of a poor response cannot therefore be explained by under dosing, but is based on poor follicle number, expressed by the predictive tests female age and ORTs. For ongoing pregnancy prediction, FSH dosage did not have a significant role in the models, neither did it alter the predictive capacity of age and ORTs.

Figure 3 ROC curves of age and ORT in the prediction of poor response and ongoing pregnancy. (A) Poor response prediction based on age and ORT. The ROC curves of age or age combined with a single or more ORT are depicted. The ROC curves for ‘Age + AMH’, ‘Age + AMH + AFC’ and ‘Age + AMH + AFC + FSH’ run toward the upper left corner, indicating a good capacity to discriminate between normal and poor responders at certain cut-off levels. (B) Ongoing pregnancy prediction based on age and ORT. The ROC curves of age or age combined with one or more ORT run almost parallel to or even cross the X = Y line, indicating that the tests are useless for pregnancy prediction. AFC, antral follicle count; AMH, anti-Müllerian hormone; FSH, follicle stimulating hormone; ORT, ovarian reserve test; ROC, receiver-operating characteristic.

The clinical implications of the present findings will necessarily remain limited to the use of ORTs in predicting poor response to controlled ovarian hyperstimulation for IVF. The real clinical value of the prediction of a poor response will depend on the consequences of the prediction. So far, clinicians do not agree on what alterations in treatment regimen may be of help improving pregnancy prospects in predicted poor responders (Tarlatzis et al., 2003; Shanbhag et al., 2007; Sunkara et al., 2007). Various (pseudo)randomized controlled trials have investigated whether individualization of the FSH treatment dose results, in not necessarily a higher ovarian response but, in higher pregnancy chances in poor responders (Harrison et al., 2001; Popovic-Todorovic et al., 2003a; Klinkert et al., 2005a; Lekamge et al., 2008; Olivennes et al., 2009). Only one study reported a dosing algorithm that would increase pregnancy chances in poor responders, while others were unable to reproduce these effects (Popovic-Todorovic et al., 2003a). For optimization of the ovarian response, two studies have shown that with an individual dose, the response could be optimized and fewer patients would have a poor response (Popovic-
Todorovic et al., 2003a; Olivennes et al., 2009). This could have consequences for the treatment efficacy and costs. Future large, well-designed randomized controlled trials are necessary to identify the best treatment option for poor responders. At present, due to the low accuracy of ORTs in pregnancy prediction, exclusion of patients other than on the basis of female age is not to be supported.

In conclusion, this IPD meta-analysis demonstrates that the ORTs, AFC and AMH are highly capable of forecasting a poor responder to ovarian hyperstimulation for IVF, even without using female age. The clinical applicability of ORT-based dose adaptation on efficacy and costs remains to be demonstrated. Even more importantly, correctly identifying patients with a very poor prognosis for success in ART will not be improved by any currently known ORT. In the field of patient selection prior to ART, female age therefore remains the most important, though modestly effective, tool.

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

Authors’ roles

Funding
This project was funded by CVZ (Assembly of health insurances in the Netherlands). However, no funding bodies had any role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of interests
Prof. F.J.M. Broekmans is a member of the external advisory board for Ferring Pharmaceuticals, Hoofddorp, the Netherlands. He receives no monetary compensation. All other authors have no potential conflict of interests.

Appendix
IMPORT study group – full list of authors:
S.L. Broer, MD, PhD. S.L.Broer-2@umcutrecht.nl, University Medical Center Utrecht, Division Woman and Baby, Department or Reproductive Medicine, Utrecht, Netherlands; J. van Disselordorp, MD, PhD. J.vandisselordorp-2@umcutrecht.nl, University Medical Center Utrecht, Division Woman and Baby, Department or Reproductive Medicine, Utrecht, Netherlands; K.A. Broeze, MD, K.A. Broeze@amc.uva.nl, Center for Reproductive Medicine, Department of Obstetrics and Gynaecology, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; M. Dolleman, MD, M.Dolleman-2@umcutrecht.nl, University Medical Center Utrecht, Division Woman and Baby, Department or Reproductive Medicine, Utrecht, Netherlands; B.C. Opmeer, PhD. B.C.Opmeer@amc.uva.nl, Academic Medical Center, Department of Clinical Epidemiology and Biostatistics Netherlands; R.A. Anderson, Richard. anderson@ed.ac.uk, Centre for Reproductive Biology, Reproductive and Developmental Sciences Edinburgh, United Kingdom; M. Ashrafi, ashrafilm@royaninstitute.org, Royan Institute Tehran, Islamic Republic of Iran; L. Bancsi, L.Bancsi@rjnsate.nl, University Medical Center Utrecht, Division Woman and Baby, Department or Reproductive Medicine, Utrecht, Netherlands; E. Carropp, ecarropp@teseo.it, ASL Bari, IVF Unit Bari, Italy; A. Copperman, acopperman@maryland.com, Reproductive Medicine Associates of New York New York, United States; T. Ebner, thomas.ebner@gespag.at, IVF Unit, Women’s General Hospital Linz Linz, Austria; M. Eldar Geva, gevatszmc.org.il, Shaare-Zedek Medical Center, IVF Unit Israel; M. Erdem, erdemem@yahoo.com, Gazi University School of Medicine, Department of Obstetrics and Gynecology Ankara, Turkey; E.M. Greenblatt, egreenblatt@msiai.on.ca, Mount Sinai Center for Fertility and Reproductive Health, Department of Reproductive Endocrinology and Infertility Toronto, Ontario, Canada; K. Jayaparakas, k.jayaparakas@nottingham.ac.uk, Academic Division of Reproductive Medicine and Surgery, School of Human Development, University of Nottingham, Nottingham, UK; Raine Fenning, Nick.Fenning@nottingham.ac.uk, Academic Division of Reproductive Medicine and Surgery, School of Human Development, University of Nottingham, Nottingham, UK; E.R. Klinkert, erklinkert@planet.nl, University Medical Center Utrecht, Division Woman and Baby, Department or Reproductive Medicine, Utrecht, Netherlands; J. Kwee, j.kwee@slaz.nl, VU University Medical Center-Department of Obstetrics/Gynaecology, Amsterdam, The Netherlands; C.B. Lambalk, cb.lambalk@vumc.nl, VU University Medical Center Department of Obstetrics/Gynaecology, Amsterdam, The Netherlands; A. La Marca, antalammarca@libero.it, University of Modena and Reggio Emilia, Mother-Infant Division, Section of Obstetrics and Gynecology Modena, Italy; M. Mclveen, myvanwy@mcilveen.com.au, University of Sydney, Department of Obstetrics and Gynaecology Sydney, New South Wales, Australia; L.T. Merce, lmerce@sego.es, Ruber International Hospital, Department of Obstetrics and Gynaecology Madrid, Spain; S. Mututkrisyna, s.mututkrisyna@ucc.ie, s.mututkrisyna@ucu.ac, Royal Free and University College Medical School, Department of Obstetrics and Gynaecology, United Kingdom; S.M. Nelson, s.nelson@clinmed.gla.ac.uk, University of Glasgow, Reproductive & Maternal Medicine Glasgow, United Kingdom; H.Y. Ng, nguye@ku.hk, University of Hong Kong, Department of Obstetrics and Gynaecology Hong Kong, Hong Kong; B. Popovic-Todorovic, drriba@yahoo.com, Copenhagen University Hospital, Rigshospitalet, The Fertility Clinic Copenhagen, Denmark; J.M.J. Smeenk, j.smeenk@elisabeth.nl, University Hospital Nijmegen, Department of Obstetrics and Gynaecology, Nijmegen, Netherlands; C. Tomás, cttomas@avaclinic.com, AVA Clinic, Fertility Center Tampere, Finland; P.J.Q. van der Linden, p.j.q.vanderlinden@dz.nl, Deventer Hospital, Department of Obstetrics and Gynaecology Deventer, Netherlands; I.A. van Rooij, iajvanrooij@gmail.com, Twee Steden Ziekenhuis, Obstetrics and Gynaecology Tilburg, Netherlands; I.K. Vladimirov, vladimirov@doctor.bg, Medical University Sofia, Obstetrics and Gynecology Sofia, Bulgaria; P.B. Bossuyt, p.m.bossuyt@amc.uva.nl,

Downloaded from https://academic.oup.com/humupd/article-abstract/19/1/26/629243 by guest on 28 February 2019
References


