AMH and AFC as predictors of excessive response in controlled ovarian hyperstimulation: a meta-analysis

S.L. Broer 1,*, M. Dölleman 1, B.C. Opmeer 2, B.C. Fauser 1, B.W. Mol 3, and F.J.M. Broekmans 1

1Department of Reproductive Medicine and Gynecology, University Medical Center, room F05.126, P.O. Box 85500, 3508 GA Utrecht, The Netherlands 2Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, The Netherlands 3Department of Obstetrics and Gynecology, Academic Medical Center, Amsterdam, The Netherlands

*Correspondence address. E-mail: s.l.broer@umcutrecht.nl
Submitted on February 18, 2010; resubmitted on June 22, 2010; accepted on June 30, 2010

TABLE OF CONTENTS
- Background
- Methods
- Results
  - Systematic review
  - Accuracy of AMH in excessive response prediction
  - Accuracy of AFC in excessive response prediction
  - Clinical value
  - Comparison of AMH with AFC
- Discussion
  - Main findings
  - Implications for clinical practice
  - Limitations
  - Future research
  - Summary

BACKGROUND: Anti-Mullerian hormone (AMH) is a marker of ovarian reserve status and represents a good predictor of ovarian response to ovarian hyperstimulation. The aim of this study was to assess the accuracy of AMH and antral follicle count (AFC) as predictors of an excessive response in IVF/ICSI treatment.

METHODS: A systematic review and meta-analysis of the existing literature was performed. Studies were included if 2 × 2 tables for the outcome excessive response in IVF patients in relation to AMH/AFC could be constructed. Using a bivariate meta-analytic model, both summary point estimates for sensitivity and specificity were calculated, as well as summary ROC curves. Clinical value was analysed by calculating post-test probabilities of excessive response at optimal cut-off levels, as well as the corresponding abnormal test rates.

RESULTS: Nine studies reporting on AMH and five reporting on AFC were found. Summary estimates of sensitivity and specificity for AMH were 82 and 76%, respectively, and 82 and 80%, respectively, for AFC. Comparison of the summary estimates and ROC curves for AMH and AFC showed no statistical difference. Abnormal test rates for AMH and AFC amounted to ~14 and 16%, respectively, at cut-off levels where test performance is optimal [likelihood ratio for a positive result (LR+) > 8], with a post-test probability of ±70%.
**Background**

In *in vitro* fertilization (IVF) treatment, excessive response to FSH stimulation introduces the risk of abdominal discomfort, painful follicle aspirations and cycle cancellations (Delvigne and Rozenberg, 2002). An excessive response will also typically introduce the risk of ovarian hyperstimulation syndrome (OHSS), a potentially life-threatening condition (Fauser et al., 2008). Excessive response to ovarian stimulation will generate many oocytes for the laboratory that will not unequivocally lead to a full range of good quality embryos (Baart et al., 2007; Heijnen et al., 2007; Verberg et al., 2009). In addition, chances of pregnancy may decrease (van der Gaast et al., 2006). In view of these drawbacks, elimination of exaggerated ovarian response in stimulation protocols will improve safety, success and costs of assisted reproduction technology (ART) programs.

For primary preventive management to be developed, the reliability of tools for prediction of ovarian response needs to be assessed first. Ovarian response prediction is mainly based on ovarian reserve tests like the antral follicle count (AFC) and Anti-Müllerian hormone (AMH) (Broekmans et al., 2006; Broer et al., 2009). The AFC comprises the number of 2–5 or 2–10 mm diameter follicles measured in the ovaries at the start of the menstrual cycle (de Carvalho et al., 2008) and is highly correlated to the number of oocytes retrieved at pick up (Kwee et al., 2007; Broer et al., 2009). AMH has been implicated as the most valuable marker of ovarian reserve as serum concentrations correlate highly with baseline AFC and the number of oocytes retrieved at aspiration (van Rooij et al., 2002; Eldar-Geva et al., 2005; Tremellen et al., 2005; Nakhuda et al., 2006, 2007; Riggs et al., 2008; Nardo et al., 2009). The aim of the present systematic literature review was to assess the true accuracy of AMH and AFC as prognostic indicators for the prediction of an excessive response after IVF/ICSI treatment.

**Methods**

**Search and selection strategy**

The literature was searched for studies that addressed the capacity of AFC or AMH as prognostic indicators of excessive ovarian response after controlled ovarian hyperstimulation in an IVF or ICSI treatment. No pre-set definition of excessive ovarian response was used. Excessive ovarian response definition included oocytes at retrieval above a certain threshold, estrogen-level above a certain threshold, the development of OHSS or cycle cancellation due to a high response, or combinations of these. Also, any cut-off or set of cut-offs for an abnormal AMH or AFC were included in this review.

A systematic search in Medline was carried out using the keywords ‘*in vitro fertilization’, ‘*in vitro fertilization’, ‘*assisted’, ‘*intracytoplasmic’, ‘*intra-cytoplasmic’ in combination with ‘anti-Mullerian hormone’, ‘Müllerean inhibiting factor’, ‘Mullerian inhibiting substance’ or ‘antral follicle count’. A period of all the years through November 2009 was covered by the search. The abstracts of all studies identified were read by one researcher (M.D.). Any article that could possibly be of value for the association between AMH and AFC and the IVF outcome excessive ovarian response was preselected. In the next step, two researchers (M.D. and S.L.B.) carefully read and judged all preselected articles independently. If it was judged possible to construct 2 × 2 tables from the data presented in the paper, the study was selected for final inclusion and analysis in this review. In a 2 × 2 or contingency table, the true positive, true negative, false positive and false negative test results at a certain cut-off are displayed. In the event of any disagreement between the two authors, the opinion of a third researcher (F.J.M.B.) was final.

The authors of studies that reported on the ovarian reserve test result in relation to IVF outcome without the possibility of constructing 2 × 2 tables were contacted by email and asked to provide the necessary data for the construction of such a table. If adequate data were obtained in this way, the study was added to the selection. In every selected study, the reference list was scanned to identify studies that could possibly be included in the selection and these were then processed as described.

Each selected study was further scored by the researchers (M.D and S.L.B.) regarding the following study quality characteristics: (i) patient sampling (consecutive or other); (ii) data collection method (prospective or retrospective); (iii) study design (cohort or case–control); (iv) blinding (present or absent); (v) selection bias, i.e. exclusion of cases based on criteria that affect the ability to generalize the findings of the study, for instance women with elevated basal FSH or women over 38 years of age (present or absent); (vi) verification bias, i.e. the use of results of the test under study in adapting the treatment protocol in order to prevent the predicted outcome, for instance poor ovarian response (present or absent); (vii) analysis upon one or multiple cycles per couple; and (viii) stimulation protocol (GnRH agonist or GnRH antagonist). Also, data on the cut-off levels used were recorded, as was the assay used for AMH measurement and whether AFC was measured in 2–5 or 2–10 mm follicles.

Because this review used only published data from the literature, no approval from our institutional review board was required.

**Analysis**

First, 2 × 2 tables were constructed from which sensitivity and specificity were calculated. Sensitivity-specificity points were displayed in the receiver operating characteristics (ROC) space (1-specificity versus sensitivity). Combinations of sensitivity and 1-specificity are indicative of the test accuracy, where studies reporting high accuracy for both sensitivity and specificity are located in the upper-left corner of the ROC space, and poor test results are located close to the x = y line.

A meta-analysis was performed using a bivariate regression model (Reitsma et al., 2005). In short, this bivariate model preserves the two-dimensional nature of prognostic data in a single model, rather than using a single outcome measure for each study such as the diagnostic odds ratio. The bivariate model simultaneously estimates sensitivity and specificity, and incorporates the negative correlation that may exist between sensitivity and specificity within studies, owing to possible implicit differences in the applied threshold between studies. When necessary, the bivariate model uses a random approach for both sensitivity and specificity, allowing for heterogeneity beyond chance due to clinical or

**Conclusions:** Both AMH and AFC are accurate predictors of excessive response to ovarian hyperstimulation. Moreover, both tests appear to have clinical value. This opens ways to explore the potential of individualized FSH dose regimens based on ovarian reserve testing.

**Key words:** Anti-müllerian hormone / antral follicle count / IVF / ovarian response / meta-analysis
methodological differences between studies. In addition, the model acknowledges the difference in precision by which sensitivity and specificity have been measured in each study. This means that studies with a larger number of women with an excessive response received more weight in the calculation of the pooled estimate of sensitivity, whereas studies with a high number of women without an excessive response were more influential in the pooling of specificity.

Sensitivity was plotted against 1-specificity (false positive rate) and pooled estimates for sensitivity and specificity were calculated and also plotted, together with the 95% confidence interval (CI) ellipse. As different studies have reported results for different thresholds to define a positive test (cut-off), we did not limit our analysis to a single threshold value, but took advantage of the fact that the model incorporates opposite effects on sensitivity and specificity when using different cut-offs. In order to account for dependent observations (observations on different cut-offs from the same study are likely to be correlated), we estimated the model in 250 stratified bootstrap samples, in which only one threshold value from each study was randomly selected. The overall estimates of sensitivity and specificity were based on the average from 250 bootstrap samples.

The results of the model were used to estimate summary ROC curves, where the increase in sensitivity and decrease in specificity reflect the shift in threshold value of the ovarian reserve test in the model. We thereby had to convert parameter estimates from the bivariate model to those in the Summary ROC model, as these are basically different statistical approaches for the same underlying model (Harbord et al., 2007). The difference between AMH and AFC in pooled sensitivity and specificity was tested by fitting the bivariate model on data for both tests, with the test included as a covariate in the model.

To assess the clinical value of both tests, post-test probabilities for the prediction of an excessive response were calculated, by using the estimated summary ROC curve and assuming an arbitrary prevalence (or pre-test probability) of 20% for an excessive response. A series of likelihood ratio ranges for an abnormal test result was then derived from several points of the estimated summary ROC curve, and at these various ranges of likelihood ratios, the post-test probabilities for both tests were computed, as well as the corresponding abnormal test rates.

All statistical analyses were done using SAS 9.1 for Windows (Proc NLMixed in the bivariate model).

**Results**

**Systematic review**

The systematic Medline search produced 170 hits. Of these, 126 articles were excluded on the basis of title and abstract. Another 30 studies were excluded on the basis of the fully read article. Finally, 14 studies were selected to be appropriate for the current meta-analysis. From those 14 studies, in four studies 2 × 2 tables could be constructed from the article itself. The remaining 10 authors were contacted and asked for the necessary data. Three studies could not be included as the authors did not reply to the email request (Seifer et al., 2002; Tremellen et al., 2005; Lekamge et al., 2007). Seven authors did respond with the appropriate data to construct 2 × 2 tables. Thus, a final number of 11 studies could be included for data extraction and meta-analysis (Ng et al., 2000; van Rooij et al., 2002; Eldar-Geva et al., 2005; Ebner et al., 2006; Kwee et al., 2008; La Marca et al., 2007; Nelson et al., 2007; Lee et al., 2008; Riggs et al., 2008; Nardo et al., 2009; Afatoonian et al., 2009). Six studies reported on the capacity of AMH to predict excessive response after IVF, two studies reported on the capacity of AFC and three studies studied both AMH and AFC (Fig. 1).

From this table, it becomes clear that all studies but one presented data for one cycle per couple and that the majority used a

---

**Figure 1** Search and selection strategy.
prospective cohort design. However, selection bias was judged to be present in all studies. This concerned the exclusion of older women or women with signs of decreased ovarian reserve, or exclusion of cases with the PCO syndrome. The definition of excessive response was not uniform. It ranged from number of oocytes retrieved over 14 up to over 21 or the development of OHSS.

### Accuracy of AMH in excessive response prediction

Sensitivities and specificities for the prediction of excessive ovarian response, as calculated from each study reporting on AMH, are summarized in Table II. A plot of sensitivity–specificity combinations in an ROC space is shown in Fig. 2. For AMH, the sensitivity varied between 40 and 95% and the specificity varied between 31 and 96%. Using the bivariate model that accounts for the heterogeneity of the studies, the summary estimates for sensitivity and specificity were calculated. The summary estimates were 82% (95% CI 52–95%) for sensitivity and 76% (95% CI 43–93%) for specificity.

Figure 2 shows the summary estimate for the overall test accuracy as calculated from the bivariate model and its 95% CI ellipse, as well as the summary ROC curve.

### Accuracy of AFC in excessive response prediction

Sensitivities and specificities for the prediction of an excessive ovarian response, as calculated from each study reporting on AFC are summarized in Table II. A plot of sensitivity–specificity combinations in an ROC space is shown in Fig. 2. For the AFC, the sensitivity varied between 20 and 94% and specificity varied between 33 and 98%. Using the bivariate model that accounts for the heterogeneity of the studies, the summary estimates for sensitivity and specificity were calculated. The summary estimate of sensitivity was 82% (95% CI 30–98%) and the summary estimate of specificity was 80% (95% CI 31–97%).

Figure 2 shows the summary estimates as calculated by the bivariate model and its 95% CI ellipse, as well as the summary ROC curve for the AFC in the prediction of an excessive response.

### Clinical value

On the basis of the summary ROC curves depicted in Fig. 2, a range of positive likelihood ratios was calculated corresponding to various sensitivity–specificity points on this ROC curves. For each of these likelihood ratio values, the pre AMH or AFC test probabilities of an excessive response were converted into post-test probabilities of an excessive response. Table III depicts a series of likelihood ratios ranges and the probability of obtaining an abnormal test result for AMH or AFC corresponding to this likelihood ratio range, as well as the post-test probability of an excessive response. At a positive likelihood ratio of at least ~8, the post-test probability of having an excessive response is close to 70%, if the pre-test probability is assumed to be ~20%. The probability of obtaining a test result for AMH or AFC with a likelihood ratio of at least ~8 is 14 and 16%, respectively.
Comparison of AMH with AFC

Comparison of summary point estimates for accuracy of the prediction of excessive response showed no statistically significant difference in the performance for AMH compared with the AFC, when sensitivity ($P = 0.87$) and specificity ($P = 0.80$) at the estimated summary cut-off point were considered. In the comparison of the estimated summary ROC curves, AFC seemed to perform slightly better than AMH, although the curves did not differ statistically. It should be noted that the summary curve for the AFC was based on fewer studies (Fig. 2).

Clinical value as outlined in Table III indicated a similar performance for AMH compared with the AFC. This is in line with the course of the ROC curves along the $y$-axis suggesting that many cases of excessive response can be identified with only a limited number of false positives. For both AMH and AFC, sensitivity can amount up to 70% with a false positive rate of 15%, and this performance level will imply a realistic number of abnormal tests ($\approx 25\%$).

Discussion

Main findings

The current meta-analysis summarizes the available evidence concerning the accuracy of AMH and the AFC in the prediction of excessive ovarian response to stimulation for IVF. It appears that both tests have a good discriminatory capacity to separate normal and excessive responders, with a definition that varies across the studies from more than 14–21 oocytes yielded. From the ROC curves (Fig. 2) it becomes clear that, currently, AMH and AFC have an equal level of accuracy in the prediction of an excessive response and that there is no statistical difference between both tests. Moreover, both AMH and AFC have clinical value, with an abnormal test rate of 14 and 16%, respectively, at cut-off levels where test performance is optimal ($LR^+ > 8$). At these cut-off levels the post-test probability of an excessive response appeared to be close to 70%.

The comparison between the AFC and AMH for their use as predictive tests for ovarian response may imply other factors than accuracy alone. For AMH as a laboratory test, measurement stability will be dealt with according to routine procedures, but routine assays may not yet be easily available. To date, two commercially available immunoassays, the Beckmann–Coulter and DSL ELISA, exist. These assays have demonstrated a very good correlation, making it possible to translate results from one to the other within the same dataset. However, there is no obvious match in absolute levels between studies; with the Beckmann–Coulter measurements reported as being approximately four to five times higher than the DSL measurements (Bersinger et al., 2007; Freour et al., 2007). Therefore, standardization of these assays is urgently needed. This situation also hampers the efforts to extract a generally applicable cut-off level for

Table II Performance of AMH and AFC in the prediction of excessive response.

<table>
<thead>
<tr>
<th>Author</th>
<th>Cycles (n)</th>
<th>Cut-off (ng/ml)</th>
<th>Abnormal test result (%)</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>LR+</th>
<th>Pre-test probability</th>
<th>Post-test probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van Rooij et al. (2002)</td>
<td>114</td>
<td>3.50</td>
<td>0.08</td>
<td>0.40</td>
<td>0.95</td>
<td>8.32</td>
<td>0.09</td>
<td>0.44</td>
</tr>
<tr>
<td>Eldar-Geva et al. (2005)</td>
<td>53</td>
<td>3.50</td>
<td>0.32</td>
<td>0.72</td>
<td>0.89</td>
<td>6.32</td>
<td>0.34</td>
<td>0.76</td>
</tr>
<tr>
<td>Ebner et al. (2006)</td>
<td>135</td>
<td>1.66</td>
<td>0.75</td>
<td>0.95</td>
<td>0.31</td>
<td>1.38</td>
<td>0.16</td>
<td>0.21</td>
</tr>
<tr>
<td>La Marca et al. (2007)</td>
<td>48</td>
<td>2.60</td>
<td>0.50</td>
<td>0.86</td>
<td>0.56</td>
<td>1.95</td>
<td>0.15</td>
<td>0.25</td>
</tr>
<tr>
<td>Nelson et al. (2007)</td>
<td>314</td>
<td>2.10</td>
<td>0.27</td>
<td>0.88</td>
<td>0.79</td>
<td>4.10</td>
<td>0.08</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>314</td>
<td>3.50</td>
<td>0.08</td>
<td>0.57</td>
<td>0.96</td>
<td>13.80</td>
<td>0.07</td>
<td>0.52</td>
</tr>
<tr>
<td>Lee et al. (2008)</td>
<td>262</td>
<td>1.99</td>
<td>0.50</td>
<td>0.90</td>
<td>0.62</td>
<td>2.38</td>
<td>0.23</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>262</td>
<td>3.36</td>
<td>0.25</td>
<td>0.62</td>
<td>0.87</td>
<td>4.64</td>
<td>0.23</td>
<td>0.38</td>
</tr>
<tr>
<td>Riggs et al. (2008)</td>
<td>123</td>
<td>1.59</td>
<td>0.49</td>
<td>0.84</td>
<td>0.67</td>
<td>2.56</td>
<td>0.31</td>
<td>0.53</td>
</tr>
<tr>
<td>Nardo et al. (2009)</td>
<td>165</td>
<td>3.50</td>
<td>0.36</td>
<td>0.88</td>
<td>0.70</td>
<td>2.90</td>
<td>0.10</td>
<td>0.24</td>
</tr>
<tr>
<td>Aflatoonian et al. (2009)</td>
<td>159</td>
<td>4.83</td>
<td>0.42</td>
<td>0.93</td>
<td>0.78</td>
<td>4.26</td>
<td>0.28</td>
<td>0.63</td>
</tr>
<tr>
<td>AFC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ng et al. (2000)</td>
<td>128</td>
<td>9</td>
<td>0.39</td>
<td>0.60</td>
<td>0.71</td>
<td>2.09</td>
<td>0.31</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>14</td>
<td>0.10</td>
<td>0.20</td>
<td>0.94</td>
<td>3.48</td>
<td>0.31</td>
<td>0.62</td>
</tr>
<tr>
<td>van Rooij et al. (2002)</td>
<td>114</td>
<td>14</td>
<td>0.42</td>
<td>0.92</td>
<td>0.63</td>
<td>2.49</td>
<td>0.10</td>
<td>0.22</td>
</tr>
<tr>
<td>Eldar-Geva et al. (2005)</td>
<td>56</td>
<td>14</td>
<td>0.78</td>
<td>0.94</td>
<td>0.33</td>
<td>1.41</td>
<td>0.40</td>
<td>0.48</td>
</tr>
<tr>
<td>Kwee et al. (2008)</td>
<td>110</td>
<td>10</td>
<td>0.38</td>
<td>0.94</td>
<td>0.71</td>
<td>3.26</td>
<td>0.15</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>110</td>
<td>12</td>
<td>0.30</td>
<td>0.88</td>
<td>0.80</td>
<td>4.33</td>
<td>0.15</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>110</td>
<td>14</td>
<td>0.21</td>
<td>0.81</td>
<td>0.89</td>
<td>7.64</td>
<td>0.15</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>110</td>
<td>16</td>
<td>0.11</td>
<td>0.50</td>
<td>0.96</td>
<td>11.75</td>
<td>0.15</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>110</td>
<td>18</td>
<td>0.06</td>
<td>0.31</td>
<td>0.98</td>
<td>14.69</td>
<td>0.15</td>
<td>0.71</td>
</tr>
<tr>
<td>Aflatoonian et al. (2009)</td>
<td>159</td>
<td>16</td>
<td>0.31</td>
<td>0.89</td>
<td>0.92</td>
<td>11.26</td>
<td>0.28</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Note: If a study reported on multiple cut-off values, data for all cut-off values are shown. LR+, likelihood ratio for a positive test result.
deciding who will be a predicted excessive responder. But, AMH is a cycle independent test (Cook et al., 2000; Hehenkamp et al., 2006; La Marca et al., 2006; Tsepelidis et al., 2007), so any measurement in the period before starting the ART cycle will be at the disposition of the clinician, making the test an ideal tool. For the AFC, standardization needs to be dealt with by the physician (Broekmans et al., 2009), implying choices on ultrasound equipment, dedicated personnel and a systematic visualization and counting process. As the intra- and between cycle stability for the AFC may be comparable to that for AMH (van Disseldorp et al., 2010), the unlimited availability of this test makes it the preferable one for the short-term.

Currently, the identification of patients at risk for excessive response is based on a variety of factors, such as age, body weight and the presence of polycystic ovaries (Kwee et al., 2007; Lee et al., 2008; Riggs et al., 2008). However, the predictive value of these factors is quite poor. Whether their addition to tests like the AFC and AMH will improve the predictive capacity in identifying excessive responders remains to be established. In fact, patients with the polycystic ovary syndrome have clearly elevated AFC’s and AMH levels (Broekmans et al., 2008; La Marca et al., 2009). As most studies have not excluded PCOS cases, these cases will add to the current analysis. In studies on PCOS cases only, AMH levels do indeed predict ovarian response to controlled hyperstimulation (Kaya et al., 2010). A limited number of studies exist on the use of multifactor prediction of the number of oocytes retrieved, using female age, basal FSH, ovarian ultrasound and smoking behaviour as predictors (Popovic-Todorovic et al., 2003; Howles et al., 2006). Validation of these prediction models in external populations has not been carried out so far. Individual patient data analyses of published literature may enable the assessment of the true value of such multivariable approach, combining patient characteristics and test results (Broeze et al., 2009).

Implications for clinical practice

Excessive response to ovarian stimulation induces the risk of the OHSS, especially in cases where exaggerated response is followed by a pregnancy. It may also cause increased patient discomfort and even reduced prospects for pregnancy. Up to 30% of IVF cycles are accompanied by complaints of mild or moderate OHSS and in 3–8% the severe form of OHSS may develop (Delvigne and Rozenberg, 2002). Once an excessive response has occurred, hCG administration could be withheld in an effort to eliminate this risk. Protective measures have also been reported for conditions in which oocyte retrieval has been allowed. Albumin administered at
that a cut-off can be used at which 75% of excessive responders will be identified, a considerable number of excessive responders will be turned into normal responders by using for instance a lower than standard dose of FSH. At the same time, the abnormal test will be falsely positive in some 15% of cases, and a lower dose may turn these cases into poor responders. Whether this ‘poor’ response may alter the prospects for pregnancy may be disputed, as mild stimulation protocols have demonstrated that in normal profile cases, a mild response does not affect outcome (Out et al., 2001; Heijnen et al., 2007; Olivennes et al., 2009; Verberg et al., 2009).

Currently, only a few studies have addressed the use of reduced dosages of FSH based on prior prediction of the ovarian response level. In the study by Popovic-Todorovic et al. (2003), individualized FSH dosing appeared not to reduce the proportion of excessive responders, although the reduced dose group produced on average two oocytes less than the standard stimulated group. In one other study, Olivennes et al., demonstrated that in predicted excessive responders, the use of FSH dosages lower than 150 IU produced mild ovarian responses without compromising pregnancy rates. A randomized comparison of standard versus individualized treatment based on the CONSORT prediction algorithm has recently been finalized and results are awaited (Olivennes et al., 2009). Such studies should not only focus on the achievement of a more homogenous ovarian response. Also cost-economic effects regarding prevention of severe OHSS and a reduction of FSH consumption will aid in rationalising ovarian stimulation protocols for IVF.

**Limitations**

Although the process of systematic literature review and meta-analysis is a practical way to generate a more powerful estimate of true effect-size with less random error than individual studies, it does come with some limitations. First of all, the heterogeneity of studies must be addressed, as it may affect the justification for pooling the data into one analysis. In the case of the present meta-analysis, heterogeneity was caused by both different study quality characteristics and slight differences in study populations. Additionally, the definition of excessive response was not uniform across studies (Table I) and varied from the use of a threshold for number of oocytes aspirated to the development of OHSS. Another limitation was the allocation, by the authors, of different cut-off values for AMH and AFC. This is problematic as it interferes with the identification of a single threshold for AMH of AFC that could be predictive of an excessive response. The solution for this problem is the construction of a ROC curve, by which the effect of different cut-offs on the sensitivity/specificity combinations will become clear and overall accuracy becomes apparent.

Many of these methodological problems may be overcome by using individual patient database meta-analysis. From such data sets, population and patient characteristics, test results, stimulation data and outcome variables can be uniformed as much as possible before applying meta-analysis. Currently, initiatives in this field have been employed (Broeze et al., 2009).

Lastly, there are some limitations that apply specifically to the method used to assess AMH levels and the AFC. The studies in this meta-analysis did not all use the same AMH assay. There is a noteworthy difference between the Beckmann—Coulter ELISA and the Diagnostic System Laboratories (DSL) ELISA leading to a wide dispersion of AMH.

---

**Table III**  Clinical value of AMH and AFC in the prediction of an excessive response.

<table>
<thead>
<tr>
<th>Prediction of an excessive response (Pre-test probability 20%)</th>
<th>Occurrence of abnormal test result in LR range (%)</th>
<th>Post-test probability of excessive response (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR range</td>
<td>AMH</td>
<td>AFC</td>
</tr>
<tr>
<td>2–3</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>4–5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6–7</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>&gt;8</td>
<td>14</td>
<td>16</td>
</tr>
</tbody>
</table>

The occurrence of both AMH and AFC results within a specified likelihood ratio (LR) range and the concomitant post-test probabilities of an excessive response are shown, given a prevalence of an excessive response of 20%. For example, at a positive likelihood ratio (LR+) of at least 6, the post-test probability is 60% if the prior chance of having an excessive response is 20%. With the cut-off levels for the test corresponding to these LR+ levels the proportion of abnormal tests is 21% for the AMH and 23% for the AFC.
concentrations (Freour et al., 2007). This compatibility problem can only be overcome by the development of an internationally standardized AMH assay (Freour et al., 2007). Similar problems arise with the use of AFC results, where either follicle sizes of 2–5 or 2–10 mm are included into the counts. Although both methods of measurement may deliver the same level of accuracy for the test, it certainly will hamper the identification of a generally applicable cut-off.

Future research
The role of ovarian reserve tests in excessive response prediction combined with simple patient characteristics could be further analysed by using large individual patient data sets. The EXPORT (individual meta-analysis of patient data for Excessive Response Prediction with Ovarian Reserve Tests) initiative may offer the opportunity to start such effort. Moreover, stimulation protocols tailored on the basis of ovarian response prediction should be analysed as to their effects on pregnancy rates, costs for medication and patient satisfaction. Only large randomized comparisons of standard treatment strategies versus individualized treatment approaches will provide the correct answers, and will enforce previous undertakings in this area (Popovic-Todorovic et al., 2003).

Summary
The current systematic review and meta-analysis demonstrates that both the AFC and AMH are capable of identifying excessive responders to ovarian stimulation for IVF. Test optimization for clinical application may be more promising for AMH.

Conflict of interest: none declared.

Funding
F.J.M.B. is a member of the external advisory board for Ferring Pharmaceuticals, Hoofddorp, The Netherlands. He receives no monetary compensation. B.C.F. has received fees and grant support from the following companies (in alphabetic order): Andromed, Ardana, Ferring, Genovum, Merck Serono, Organon, Pantheare Bioscience, PregLem, Schering, Schering Plough, Serono and Wyeth.

References
Hekenkamp WJ, Looman CVW, Themmen AP, de Jong FH, te Velde ER, Broekmans FJ. Anti-Mullerian hormone levels in the spontaneous menstrual cycle do not show substantial fluctuation. f Clin Endocrinol Metab 2006;91:4057–4063.
Kaya C, Pabuccu R, Satozyoglu H. Serum anti-mullerian hormone concentrations on day 3 of the in vitro fertilization stimulation cycle are predictive of the fertilization, implantation, and pregnancy in polycystic ovary syndrome patients.


