The role of antimullerian hormone in prediction of outcome after IVF: comparison with the antral follicle count

Simone L. Broer, B.Sc., a Ben Willem J. Mol, M.D., Ph.D., b Dave Hendriks, M.D., Ph.D., a and Frank J. M. Broekmans, M.D., Ph.D. a

a Department of Reproductive Medicine, Division of Obstetrics, Neonatology, and Gynecology, University Medical Center Utrecht; and b Department of Obstetrics and Gynecology, Academic Medical Center, Amsterdam, The Netherlands

Objective: To assess the value of antimullerian hormone (AMH) as a test to predict poor ovarian response and pregnancy occurrence after IVF and to compare it with the performance of the antral follicle count (AFC).

Design: A systematic review of existing literature and a meta-analysis were carried out. After a comprehensive search, studies were included if 2×2 tables for outcomes poor response and pregnancy in IVF patients in relation to AMH or AFC could be constructed.

Setting: Academic referral center for tertiary care.

Patient(s): Cases indicated for IVF.

Intervention(s): None.

Main Outcome Measure(s): Poor response and nonpregnancy after IVF.

Result(s): A total of 13 studies were found reporting on AMH and 17 on AFC. Because of heterogeneity among studies, calculation of a summary point estimate for sensitivity and specificity was not possible. However, for both tests summary receiver operating characteristic curves for the outcome measures poor response and nonpregnancy could be estimated and compared. The curves for the prediction of poor response indicated no significant difference between the performances of AMH and AFC. For the prediction of nonpregnancy, poor performance for both AMH and AFC was found.

Conclusion(s): In this meta-analysis it was shown that AMH has at least the same level of accuracy and clinical value for the prediction of poor response and nonpregnancy as AFC. (Fertil Steril 2008; —:—–– —. ©2008 by American Society for Reproductive Medicine.)

Key Words: Antimullerian hormone, antral follicle count, IVF, poor response, pregnancy, meta-analysis

Increasing female educational levels and participation in the labor force has resulted in a clear rise in the mean age at which women deliver their first child in Western-style societies (1). As natural fertility starts to decline after the age of 30, many women therefore will be faced with unexpected problems in becoming pregnant owing to decreased ovarian reserve (2–4). It has been shown that the rate of the ovarian reserve decline varies considerably between individual women, making it a challenge to design tests that estimate an individual’s remaining reproductive lifespan at a given age (4).

Ovarian reserve relates to both the quantity and the quality of the ovarian follicle pool. The number of primordial follicles that are left in the ovary at a given age is therefore an important indicator for ovarian reserve and dictates reproductive events, such as age at menopause (4). Although direct measurement of the primordial follicle pool is impossible, it has been shown that the number of antral follicles in the ovaries is proportionally related to the size of primordial follicle stock from which they were recruited (5). Therefore, the antral follicle count (AFC) is believed to represent the quantitative aspect of ovarian aging (6). Unfortunately, markers that may directly reflect oocyte quality are clearly lacking at the moment. Consequently, the age-related decrease in fertility cannot be determined through a direct test. Only through measurement of the quantity of the oocytes can information on the quality aspects of ovarian reserve be obtained (7, 8).

Ovarian response to ovarian hyperstimulation in IVF is another way in which the quantitative ovarian reserve may come to expression. Although poor response may be considered to be a sign of diminished ovarian reserve, it may also be caused by other factors, such as underdosing in obesity or in certain FSH-receptor polymorphisms (9). Assessment of the true nature of a poor ovarian response may help to direct the
management of the patient (7, 10, 11). Additionally, correct identification of poor responders, especially in older patients, before entering an IVF program is important, because it could help in proper management regarding gonadotropin dosing and denial of treatment. For this purpose, the tests of choice are currently the AFC or basal FSH, as was shown in a comparative review (12).

Antimullerian hormone (AMH), a member of the transforming growth factor β family, is produced in the granulosa cells (13). The highest level of AMH expression is present in granulosa cells of secondary, preantral, and small antral follicles up to 6 mm in diameter (14), whereas in follicles growing into dominance this expression ceases (15, 16). Antimullerian hormone is barely detectable at birth and reaches the highest values after puberty, then decreases progressively with age and becomes undetectable at menopause (17, 18). Serum AMH levels have been shown to strongly correlate with the number of antral follicles (19, 20) and have appeared to be cycle independent (21, 22). From several studies, AMH has emerged as a predictor of ovarian response to hyperstimulation (23, 24) and possibly even of the chance of becoming pregnant after IVF (25).

The aim of the present systematic review was to assess the true accuracy of AMH as a prognostic factor for the outcome of IVF/intracytoplasmic sperm injection (ICSI) treatment compared with AFC, which has been shown to be the best predictor of poor response after IVF (12).

MATERIALS AND METHODS

In the present review, studies were enrolled that addressed the evaluation of AFC and AMH as predictors of the outcomes poor response and pregnancy after IVF or ICSI treatment. No preset criteria for the definition of poor ovarian response or pregnancy were used. Poor ovarian response definition included cycle cancelation, number of dominant follicles at ultrasound, oocytes at retrieval below a certain threshold, or combinations of these. Pregnancy definition included both clinical and ongoing pregnancy. Also, any cut-off or set of cut-offs for an abnormal test result were included in the review and analysis.

A systematic search of Medline was carried out using the keywords “in vitro fertilization” or “in vitro fertilisation” or “assisted,” “intracytoplasmatic,” or “intracytoplasmic” in combination with “antimullerian hormone,” “mullerian-inhibiting substance,” or “mullerian-inhibiting factor.” A period including all years through 2006 was covered by the search. The abstracts of all the studies identified were read by one researcher (S.B.). Any article that could possibly be of value for the association between AMH and the IVF outcome poor ovarian response or pregnancy was preselected. In the next step, two researchers (S.B. and D.H.) carefully read and judged all preselected articles independently. If it was judged possible to construct 2 × 2 tables, where test result at a certain cut-off was related to the outcome parameters poor response and/or pregnancy, the study was selected for final recordings and analysis. In the event of any disagreement between the two researchers, the opinion of a third (F.B.) was final. The authors of studies that related test result to IVF outcome without the possibility of constructing a 2 × 2 table were contacted by e-mail and asked to provide the necessary data for the construction of such a table. If adequate data were obtained 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 then processed as described.

Each selected study was further scored by the researchers S.B. and D.H. on the following study quality characteristics: 1) sampling (consecutive vs. other); 2) data collection (prospective vs. retrospective); 3) study design (cohort vs. case-control study); 4) blinding (present or absent); 5) selection bias; 6) verification bias; 7) analysis on one or multiple cycles per couple; and 8) stimulation (GnRH-agonist or GnRH-antagonist). Also, data on the cut-off levels used were recorded, as was the assay used for AMH measurement.

For the comparison of AMH and AFC, we updated the recently published meta-analyses (8, 12) on the performance of AFC. The period to be covered by the systematic search for studies reporting on AFC in the prediction of poor response and nonpregnancy after IVF was extended through December 2006 (8, 12). The same basic series of keywords was used as listed above, in combination with “antral follicle count” or “antral follicle number.” If by this search new studies were found and judged suitable for processing according to the described procedure, they were added to the already analyzed AFC studies. If a study on both AMH and the AFC was located by any of the search strategies, this study was used for both review groups.

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

Analysis

First, for each study finally included, we calculated sensitivity and specificity from the 2 × 2 tables. Sensitivity-specificity points were plotted in a receiver operating characteristic (ROC) curve. Homogeneity of the sensitivity and specificity points was tested by means of the chi-squared test. A summary point estimate of sensitivity-specificity points and 95% confidence interval (CI) was calculated if homogeneity for both parameters could not be rejected. In case of heterogeneity for one or both parameters, logistic regression was used to evaluate whether the study characteristics were associated with the discriminatory capacity. If one of the study characteristics was found to have a statistically significant impact on the performance of the test, further analysis was performed in subgroups of patients. If not, it was explored whether the differences in sensitivity and specificity combinations were the result of the use of different threshold levels. For that purpose, a Spearman correlation coefficient was calculated for the association between sensitivity and specificity.
In case of a negative correlation as defined by a correlation coefficient of −0.5 or less, a summary ROC curve was estimated, using a random effects regression model (26–28) and assuming that studies were heterogenous because of the use of different threshold levels. The same procedures were followed for studies on AFC from the updated search.

The constructed summary ROC curves for AMH and AFC were tested for statistically significant differences with a linear regression model, similar to the model used to evaluate the impact of study characteristics.

RESULTS

Systematic Review

The systematic Medline search produced 742 hits, from which we selected 24 studies based on the abstracts. We were able to create 2 × 2 tables from nine studies. We contacted the authors from the remaining studies, four of whom provided us with the necessary data to construct 2 × 2 tables.

Through this search and selection strategy, a final number of 13 studies reporting on the capacity of AMH to predict ovarian response and/or nonpregnancy after IVF and considered suitable for data extraction and meta-analysis were identified (20, 29–40). Five studies reported on both poor response and pregnancy, one study on pregnancy alone, and seven studies on poor response alone. The characteristics of the included studies are listed in Table 1. From this table it was shown that all studies presented data for one cycle per couple and that the majority used a prospective cohort design. Also, definitions for poor response were quite uniform. However, selection bias was judged to be present in quite a number of studies.

For the AFC, the updated systematic search and selection revealed no additional studies eligible for analysis. Consequently, a total of 17 studies on AFC were available.

Accuracy of Poor Response Prediction

Sensitivities and specificities for the prediction of poor ovarian response, as calculated from each study reporting on AMH, are summarized in Table 2. The sensitivity varied between 40% and 91% and the specificity between 64% and 100%. Homogeneity for both sensitivity and specificity had to be rejected (P value for the χ² test for sensitivity and specificity: .04 and .001, respectively). For this reason, the calculation of a single summary point estimate for sensitivity and specificity was not meaningful.

Logistic regression analysis showed that none of the study characteristics recorded had a statistically significant impact on the reported predictive performance of AMH. For example, whether the design of the study was retrospective or prospective, no influence was made on the prognostic capacity of AMH as estimated by the studies. A plot of sensitivity-specificity points in an ROC space is shown in Figure 1. The Spearman correlation coefficient for sensitivity and specificity was −0.31, which was judged to be sufficient to estimate a summary ROC curve (Fig. 1).

Accuracy of Nonpregnancy Prediction

For the prediction of nonpregnancy, the sensitivities and specificities of each study are summarized in Table 2. Similarly as for ovarian response, homogeneity for sensitivity had to be rejected. However, specificity appeared to be homogeneous (χ² test: P = .11). Sensitivity varied between 19% and 66%, whereas specificity varied between 55% and 89%. As for the estimation of one summary point for sensitivity and specificity, statistical homogeneity for both test parameters was required. Consequently, this solution was abandoned. A plot of sensitivity-specificity points in an ROC space is shown in Figure 1.

The Spearman correlation coefficient for sensitivity and specificity was −0.71 for the prediction of nonpregnancy, which was judged to be sufficient to estimate a summary ROC curve (Fig. 1).

Clinical Value

Based on the summary ROC curves depicted in Figure 1, a range of positive likelihood ratios was calculated corresponding to various sensitivity-specificity points on this ROC curve. For each of these likelihood ratio values the pre–AMH test probability of poor response or nonpregnancy (set at 20% and 80%, respectively) were converted into a post–AMH test probability. Table 3 depicts a series of likelihood ranges and the probability of obtaining an AMH test corresponding to this likelihood ratio range, as well as the posttest probability of poor response and nonpregnancy. At a maximum positive likelihood ratio of ∼8, the post–AMH test probability of poor response will approximate 65% if the pre–AMH test probability is assumed to be as high as 20%. The probability of obtaining a test result for AMH with a likelihood ratio of ∼8 is high enough to consider AMH to be a clinically valuable test for poor response prediction.

For prediction of nonpregnancy, the extremely low AMH cut-off level that is necessary to obtain a moderate positive likelihood ratio of ∼5, leading to a posttest pregnancy rate of less than 5% based on a pretest rate of 20%, occurs only in an extremely limited number of patients (Table 3). The summary ROC curve runs not far from the line of equality, indicating that most of the ROC curve is uninformative (likelihood ratio ∼1).

Comparison of AMH with AFC

In the analysis of the 17 available studies, sensitivity and specificity for AFC in the prediction of poor response and nonpregnancy both showed heterogeneity. After excluding the necessity of subgroup analysis from the study characteristic analysis, the Spearman correlation coefficients between sensitivity and specificity for both poor response and nonpregnancy were judged to be sufficient to estimate summary ROC curves (−0.63 and −0.67, respectively). These curves are drawn in Figure 1 and show quite a high accuracy for the prediction of poor ovarian response but very limited accuracy for nonpregnancy prediction.
## TABLE 1

Characteristics of the 13 included studies for antimullerian hormone (AMH).

<table>
<thead>
<tr>
<th>Author</th>
<th>Consecutive</th>
<th>One cycle per couple</th>
<th>Data per cycle</th>
<th>Cohort/ case-control</th>
<th>Prospective/ retrospective</th>
<th>Blinding</th>
<th>Selection bias</th>
<th>Verification bias</th>
<th>Agonist/ antagonist</th>
<th>Definition of poor response; pregnancy</th>
<th>AMH assay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ebner 2006 (37)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Both</td>
<td>&lt;4 oocytes; clinical &lt;4 oocytes or cancelation; ongoing</td>
<td>Beckman-Coulter Immunotech-Coulter</td>
</tr>
<tr>
<td>Muttukrishna 2004 (30)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Both</td>
<td>&lt;4 oocytes</td>
<td>Immunotech-Coulter</td>
</tr>
<tr>
<td>Van Rooij 2002 (20)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Agonist</td>
<td>&lt;4 oocytes</td>
<td>Immunotech-Coulter</td>
</tr>
<tr>
<td>Penarrubia 2005 (31)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Case-control</td>
<td>Retrospective</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Agonist</td>
<td>Cancelation; clinical ≤ 4 oocytes</td>
<td>Immunotech-Coulter</td>
</tr>
<tr>
<td>Tremellen 2005 (32)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Agonist</td>
<td>&lt;5 oocytes</td>
<td>Immunotech-Coulter</td>
</tr>
<tr>
<td>Ficicioglu 2006 (35)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Agonist</td>
<td>&lt;4 oocytes or cancelation ≤ 4 oocytes</td>
<td>Immunotech-Beckman Coulter</td>
</tr>
<tr>
<td>La Marca 2006 (22)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Agonist</td>
<td>&lt;4 oocytes</td>
<td>Beckman Coulter</td>
</tr>
<tr>
<td>McIlvane 2007 (34)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Agonist</td>
<td>&lt;6 oocytes</td>
<td>Immunotech-Beckman Coulter</td>
</tr>
<tr>
<td>Kwee 2007 (40)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Agonist</td>
<td>&lt;6 oocytes</td>
<td>Beckman Coulter</td>
</tr>
<tr>
<td>Muttukrishna 2005 (29)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Retrospective</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Agonist</td>
<td>≤ 4 oocytes</td>
<td>Immunotech-Coulter</td>
</tr>
<tr>
<td>Eldar-Geva 2005 (33)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Agonist</td>
<td>Ongoing</td>
<td>Immunotech-Coulter</td>
</tr>
<tr>
<td>Freour 2007 (39)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Not clear</td>
<td>&lt;6 oocytes</td>
<td>Beckman Coulter/ DSL</td>
</tr>
<tr>
<td>Smeenk 2007 (38)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Cohort</td>
<td>Prospective</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Agonist</td>
<td>≤ 4 oocytes; ongoing</td>
<td>Immunotech-Coulter</td>
</tr>
</tbody>
</table>

## TABLE 2

**Performance of antimullerian hormone (AMH) in the prediction of poor response and nonpregnancy in IVF patients and shift from pretest to posttest probability of poor response and nonpregnancy for patients with an abnormal test result.**

<table>
<thead>
<tr>
<th>Author</th>
<th>Cycles (n)</th>
<th>AMH cut-off value</th>
<th>Prediction characteristics</th>
<th>Pretest probability (%)</th>
<th>Posttest probability (%)</th>
<th>Proportion with abnormal test (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sens</td>
<td>Spec</td>
<td>LR+</td>
<td></td>
</tr>
<tr>
<td>Poor response</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Van Rooij 2002 (20)</td>
<td>119</td>
<td>0.1 μg/L</td>
<td>0.49</td>
<td>0.94</td>
<td>8.2</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>119</td>
<td>0.2 μg/L</td>
<td>0.54</td>
<td>0.90</td>
<td>5.4</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>119</td>
<td>0.3 μg/L</td>
<td>0.60</td>
<td>0.89</td>
<td>5.5</td>
<td>29</td>
</tr>
<tr>
<td>Muttukrishna 2004 (30)</td>
<td>69</td>
<td>0.1 ng/mL</td>
<td>0.76</td>
<td>0.88</td>
<td>6.3</td>
<td>25</td>
</tr>
<tr>
<td>Muttukrishna 2005 (29)</td>
<td>108</td>
<td>0.2 ng/mL</td>
<td>0.87</td>
<td>0.64</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>Penarrubia 2005 (31)</td>
<td>80</td>
<td>4.9 pmol/l</td>
<td>0.40</td>
<td>0.92</td>
<td>5.0</td>
<td>25</td>
</tr>
<tr>
<td>Ebner 2006 (37)</td>
<td>141</td>
<td>1.66 ng/mL</td>
<td>0.69</td>
<td>0.86</td>
<td>4.9</td>
<td>26</td>
</tr>
<tr>
<td>Ficicioglu 2006 (35)</td>
<td>44</td>
<td>0.25 pg/mL</td>
<td>0.91</td>
<td>0.91</td>
<td>10.1</td>
<td>25</td>
</tr>
<tr>
<td>La Marca 2006 (22)</td>
<td>48</td>
<td>0.5 ng/mL</td>
<td>0.85</td>
<td>0.82</td>
<td>4.7</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>0.75 ng/mL</td>
<td>0.80</td>
<td>0.93</td>
<td>11.4</td>
<td>25</td>
</tr>
<tr>
<td>Mcllvineen 2007 (34)</td>
<td>84</td>
<td>1.25 ng/mL</td>
<td>0.58</td>
<td>0.75</td>
<td>2.3</td>
<td>57</td>
</tr>
<tr>
<td>Smeenk 2007 (38)</td>
<td>80</td>
<td>1.4 μg/L</td>
<td>0.62</td>
<td>0.73</td>
<td>2.3</td>
<td>16</td>
</tr>
<tr>
<td>Tremellen 2005 (32)</td>
<td>75</td>
<td>8.1 pmol/l</td>
<td>0.80</td>
<td>0.85</td>
<td>5.3</td>
<td>27</td>
</tr>
<tr>
<td>Freour 2007 (39)</td>
<td>69</td>
<td>1.30 μg/L</td>
<td>0.44</td>
<td>1.00</td>
<td>∞</td>
<td>13</td>
</tr>
<tr>
<td>Kwee 2007 (40)</td>
<td>104</td>
<td>0.8 μg/L</td>
<td>0.55</td>
<td>0.94</td>
<td>9.2</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>1.0 μg/L</td>
<td>0.66</td>
<td>0.94</td>
<td>11</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>1.2 μg/L</td>
<td>0.69</td>
<td>0.88</td>
<td>5.8</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>1.4 μg/L</td>
<td>0.76</td>
<td>0.86</td>
<td>5.4</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>1.6 μg/L</td>
<td>0.79</td>
<td>0.78</td>
<td>3.6</td>
<td>27</td>
</tr>
<tr>
<td>Pregnancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Van Rooij 2002 (20)</td>
<td>106</td>
<td>0.1 ng/mL</td>
<td>0.22</td>
<td>0.89</td>
<td>2.0</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>106</td>
<td>0.2 ng/mL</td>
<td>0.27</td>
<td>0.85</td>
<td>1.8</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>106</td>
<td>0.3 ng/mL</td>
<td>0.28</td>
<td>0.81</td>
<td>1.5</td>
<td>75</td>
</tr>
<tr>
<td>Eldar-Geva 2005 (33)</td>
<td>56</td>
<td>18 pmol/l</td>
<td>0.67</td>
<td>0.69</td>
<td>2.2</td>
<td>54</td>
</tr>
<tr>
<td>Penarrubia 2005 (31)</td>
<td>80</td>
<td>Not stated</td>
<td>0.62</td>
<td>0.55</td>
<td>1.4</td>
<td>66</td>
</tr>
<tr>
<td>Ebner 2006 (37)</td>
<td>132</td>
<td>1.66 ng/mL</td>
<td>0.19</td>
<td>0.69</td>
<td>0.6</td>
<td>51</td>
</tr>
<tr>
<td>Smeenk 2007 (38)</td>
<td>80</td>
<td>1.4 μg/L</td>
<td>0.38</td>
<td>0.73</td>
<td>1.4</td>
<td>50</td>
</tr>
<tr>
<td>Kwee 2007 (40)</td>
<td>104</td>
<td>1.4 μg/L</td>
<td>0.34</td>
<td>0.79</td>
<td>1.6</td>
<td>77</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; Sens = sensitivity; Spec = specificity.

Comparison of the estimated summary ROC curves for the prediction of poor response showed no significant improvement in the performance for AMH compared with AFC \((P=.73)\). The overall accuracy for predicting nonpregnancy was poor for both tests. There was no significant difference between the ROC curves for the prediction of nonpregnancy between both tests \((P=.67)\).

Clinical value as outlined in Table 3 indicates a slightly better performance for basal AMH compared with AFC. Especially the course of the ROC curve along the y-axis suggests that many cases of poor response can be identified with only a limited number of false positives. If more false positives are accepted, sensitivity can amount up to 70% with only a false positive rate of 10%, and this test performance will imply a realistic number of abnormal tests.

**DISCUSSION**

**Main Findings**

This meta-analysis summarizes the available evidence on the accuracy of AMH compared with AFC in the prediction of poor ovarian response after stimulation for IVF. The ROC curves do not suggest a clearly better predictive ability for AMH than for AFC, and the difference was not statistically significant \((P=.73)\). This implies that the best poor response predictor to date, AFC \((12)\), has obtained company from a test that may have some crucial advantages. Application of this test does not need to be carried out on a specific day of the cycle, because AMH levels have been shown to fluctuate only marginally and prediction by samples of any cycle day will be equally accurate \((21, 36, 41)\). Blood sampling often is part of preparation for IVF treatment, and therefore extra venapuncturing will not be necessary. Currently, the availability of the AMH assay may present some problems but surely this test system will soon become part of one of the large automated platforms, with inherent validity checks and limited assay variation. In contrast, AFC necessitates skilled ultrasound operators who carefully identify, measure, and count ovarian follicles. Although observer bias may be limited technically \((42, 43)\), a new source of bias may arise from the fact that the ultrasound operator is aware of the cut-off for test judgement and may become influenced by the consequences of the test for the treatment of the couple. Such test inflation has recently been suggested from a study in older IVF patients who were allowed or refused IVF treatment on the basis of this test \((11)\). Also, AFC has to be carried out in the early follicular phase of the cycle, although variation of counts across the cycle may be very modest \((44)\).

The performance for nonpregnancy prediction is clearly poor for both AMH and the AFC. This comes as no surprise, because AMH, like AFC, is strongly thought to represent only the size of the cohort of FSH-sensitive follicles continuously present in the ovaries. Response to ovarian hyperstimulation will be directly linked to this cohort size \((45)\). The relation between quantity and oocyte and embryo quality is much less clear. Indeed, the chance of pregnancy after IVF depends on many more factors than the cohort size alone, such as embryo quality, transfer technique, and endometrial receptivity \((46)\). Also, over the past decades, not a single ovarian reserve test has been evaluated in a series of subsequent IVF cycles. It is likely that only by studying several consecutive cycles can a true representation of a woman’s remaining reproductive capacity be obtained. Only one study has demonstrated a certain predictive value regarding the occurrence of pregnancy in a selected group of cases with normal FSH and AFC levels \((25)\). That study, however, could not be included in the present meta-analysis owing to the lack of data to produce a contingency table.

**Limitations**

This meta-analysis has possible weaknesses. First, across studies poor response was defined in various ways. Most definitions were based upon outcome parameters of the IVF...
treatment, such as cycle cancelation for absent or very limited follicle growth, the number of oocytes obtained, or the number of mature follicles at ultrasound. Besides the definition of poor response, the cut-off values of poor response also differed among studies; for example, the number of oocytes retrieved may vary between $<4$ and $<7$ oocytes. This may lead to heterogeneous study groups and therefore to potential difficulties in pooling data. However, because studies appeared quite homogeneous regarding the quality characteristics analyzed, the spread across the ROC diagram indicates that cut-off values and definitions for the outcome variable used are very likely to be the cause for this variation. This is also exemplified by the fact that a summary ROC curve could be fitted to the studies.

It should be remembered that the purpose of any ovarian reserve test is the identification of women with poor ovarian reserve for their age. The present meta-analysis assessed the performance of AMH in a univariate context, independent of female age, although female age is the most important predictor in a priori prospects for IVF outcome (47). Therefore, clinical studies in which the performance of AMH is assessed in a multivariable analysis taking into account its interaction with female age are needed before the true applicability of AMH can be established.

Finally, no international assay standard for AMH measurements exists, possibly contributing to the discordance among studies and therefore making comparison between laboratories difficult (31). Also, there is a moderate intercycle and interobserver variability in AFC (42). Currently, the role of these factors cannot be separately analyzed.

**Implications for Clinical Practice**

The question is whether the fairly good predictive ability of AMH regarding the occurrence of poor response to hyperstimulation has clinical value. An ideal ovarian reserve test should identify a substantial percentage of IVF-indicated cases with a practically zero chance of becoming pregnant because of the adverse effects of diminished ovarian reserve. Those cases can be refrained from entering the program, and a quite high patient burden with only disappointing outcome can thereby be prevented. Also, high costs for only minimal results could be avoided. Accurate prediction of poor response could therefore have clinical value if the pregnancy prospects are so unfavorable that a predicted poor responder would be denied treatment.

Here we face two problems. One is the fact that poor response prediction is not fully reliable and that women with false-positive test results may incorrectly be refrained from IVF. From the ROC curve in Figure 1 it can be read that at a desired level of sensitivity of 70%–80% a false positive rate of 10%–20% can be expected. If predicted poor responders indeed have a very poor prognosis for pregnancy and should be refused treatment, extreme cut-off values would be used to prevent false positives. This would impact that only small percentages of abnormal tests will be found and many poor responders will pass unrecognized.

---

**TABLE 3**

<table>
<thead>
<tr>
<th>Predictors of poor response (pretest probability = 20%)</th>
<th>Occurrence of test results in range (%)</th>
<th>Posttest probability of poor response (%)</th>
<th>Predictors of nonpregnancy (pretest probability = 80%)</th>
<th>Occurrence of test results in range (%)</th>
<th>Posttest probability of nonpregnancy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR range</td>
<td>AMH</td>
<td>AFC</td>
<td>&lt;20</td>
<td>LR range</td>
<td>AMH</td>
</tr>
<tr>
<td>0–1</td>
<td>66</td>
<td>68</td>
<td>1–2</td>
<td>75</td>
<td>77</td>
</tr>
<tr>
<td>1–2</td>
<td>7</td>
<td>10</td>
<td>20–33</td>
<td>1–2</td>
<td>15</td>
</tr>
<tr>
<td>2–3</td>
<td>5</td>
<td>4</td>
<td>33–43</td>
<td>2–3</td>
<td>6</td>
</tr>
<tr>
<td>3–4</td>
<td>7</td>
<td>6</td>
<td>43–50</td>
<td>3–4</td>
<td>1</td>
</tr>
<tr>
<td>4–5</td>
<td>1</td>
<td>0</td>
<td>50–56</td>
<td>4–5</td>
<td>3</td>
</tr>
<tr>
<td>5–6</td>
<td>1</td>
<td>0</td>
<td>56–60</td>
<td>5–6</td>
<td>0</td>
</tr>
<tr>
<td>6–7</td>
<td>0</td>
<td>0</td>
<td>60–64</td>
<td>6–7</td>
<td>0</td>
</tr>
<tr>
<td>7–8</td>
<td>0</td>
<td>0</td>
<td>64–67</td>
<td>7–8</td>
<td>0</td>
</tr>
<tr>
<td>&gt;8</td>
<td>13</td>
<td>12</td>
<td>&gt;67</td>
<td>&gt;8</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note: For a high level of LR (i.e., ~8) the probability of producing a poor response is ~70%. The chance of obtaining such a test result at the cut-off level for AMH used would be ~13%. At the same high level of positive LR the chance of not becoming pregnant is ~97%. The probability of measuring AMH at that low cut-off, however, is close to zero.*


---

Fertility and Sterility®
Two, many poor responders achieve a pregnancy, although prospects indeed are less optimal compared with normal responders (48). Especially poor responders at a young age have a different prognosis than older poor responders (10). It is in fact the lack of a direct relation between quantity of response and quality of the oocytes that makes identification of very poor prognosis cases so difficult. In addition, the valuation of the weight of both false positive and false negative predictions should be considered. If patients are interviewed on the incorrect withholding IVF compared with incorrect performing IVF, they consider the first much worse than the second, thereby implying that currently available tests for ovarian reserve have in themselves insufficient accuracy to withhold IVF (49).

Apart from the predictive meaning for the occurrence of pregnancy after IVF, the prediction of poor ovarian response is also potentially important for individual adjustment of the dose of gonadotropins before IVF. Patients with a poor expected response are believed to benefit from a starting dose of 225 IU/day instead of 150 IU/day. A randomized trial on the subject showed that an individual dose regimen in a well defined “standard” patient population increased the proportion of appropriate ovarian responses and decreased the need for dose adjustment during controlled ovarian stimulation (50). In contrast, a randomized trial in predicted poor responders based on a prior AFC showed no benefit for response nor pregnancy rates of a stimulating dose of 300 IU compared with a dose of 150 IU (51).

In fact, the point may be raised as to whether there is any proven effective management for poor responders. In short, there are two strategies. The first method functions by using higher amounts of gonadotropins, and the second method functions on the thought that ovarian sensitivity may improve with the addition of medication. Although high doses of gonadotropins have been used by the vast majority of authors, results have been controversial and prospective randomized studies have shown little or no benefit. Adjuvant therapy with growth hormone (GH) or GH-releasing factors did not result in significant improvement. The use of corticosteroids and nitric oxide donors has shown encouraging results, but confirmation studies are lacking. Finally, natural cycle IVF has produced results which are comparable to those obtained with stimulated cycles in true poor responders. Well designed, large-scale, randomized, controlled trials are needed to assess the true efficacy of these different management strategies (52).

To date, basal FSH is the most commonly used test for ovarian reserve estimations. The accuracy and clinical value of this test was much debated in a recent review (6). In a comparison with AFC, basal FSH appeared inferior in poor response prediction (12). Therefore, based on the present results, AMH may become a test for quantitative ovarian reserve that is to be preferred over basal FSH. Apart from the cycle instability of FSH (21) compared with AMH, FSH levels may become elevated owing to other causes, such as familial dizyosity or FSH-receptor polymorphisms (9, 53, 54).

Based on the current status of ovarian reserve tests, it can be proposed that IVF may be initiated without any ovarian reserve test carried out. The response in the first cycle could then serve as a first-line test, and if it is poor it may necessitate the application of an ovarian reserve test such as AFC or AMH. If such a test would confirm the existence of a poor ovarian reserve, then the prognosis can be considered poor and further treatment refrained (55). If the test is normal, a dose adaptation may be considered to be worthwhile and continuation of treatment justified.

**Future Research**

Knowledge regarding the processes that dictate reproductive aging is still limited. We understand that follicle numbers decline with age but lack knowledge on how follicle reserve builds up in the fetal ovaries and is subsequently wasted. Thus, we cannot explain interindividual variation in this reduction process. We recognize that oocytes lose the competence to produce viable embryos with advancing age, but fail to understand the mechanisms behind this process. Two decades of research on ovarian reserve have not delivered a highly accurate endocrine or imaging test that makes a clear clinical difference in patient management. Identifying genetic markers of the processes that regulate follicle quantities and oocyte quality (56) as well as longitudinal studies on the relationship between these markers and the occurrence of menopause (57) appear to be needed to truly advance the field of assessing ovarian aging and predicting reproductive potential on an individual basis.

In summary, the present meta-analysis has shown that AMH has at least the same level of accuracy and clinical value for the prediction of poor response and nonpregnancy as AFC. Clinical applicability ultimately depends on the way abnormal test results might alter patient management.

**Acknowledgments:** The authors gratefully thank the following individuals for their cooperation in providing us with the necessary data: Prof. Dr. Ebner, IVF-Unit, Landes-, Frauen- and Kinderklinik, Linz, Austria; D. Masson, Laboratory of Clinical Chemistry, CHU Nantes, France; J. M. J. Smeenk, M.D., Department of Obstetrics and Gynecology, Radboud University Nijmegen Medical Center, The Netherlands; Dr. T. Eldar-Geva, IVF-Unit, Institute of Hormone Research and Ultrasound Unit, Department of Obstetrics and Gynecology, Shaare-Zedek Medical Center, Ben Gurion University of the Negev, Jerusalem, Israel; and J. Kwee, Division of Reproductive Endocrinology and Fertility and the IVF Centre, Department of Obstetrics and Gynecology, Vrije Universiteit Medical Centre, Amsterdam, The Netherlands.

**REFERENCES**

5. Gougeon A, Echouchard R, Thalabard JC. Age-related changes of the population of human ovarian follicles: increase in the disappearance


24. van Rooij IA, de Jong E, Broekmans FJ, Looman CW, Habbema JD, te Velde ER. High follicle-stimulating hormone levels should not necessarily lead to the exclusion of subfertile patients from treatment. Fertil Steril 2004;81:1478–85.


