Prediction of an excessive response in in vitro fertilization from patient characteristics and ovarian reserve tests and comparison in subgroups: an individual patient data meta-analysis

Simone L. Broer, M.D., Ph.D., a Madeleine Dölleman, M.D., a Jeroen van Disseldorp, M.D., Ph.D., a Kimiko A. Broeze, M.D., b Brent C. Opmeer, Ph.D., c Patrick M. M. Bossuyt, Ph.D., c Martinus J. C. Eijkemans, Ph.D., d Ben Willem Mol, M.D., Ph.D., b and Frank J. M. Broekmans, M.D., Ph.D., a on behalf of the IPD-EXPORT Study Group

a Department of Reproductive Medicine and Gynaecology, University Medical Center Utrecht; b Department of Obstetrics and Gynaecology, Academic Medical Centre, Amsterdam; c Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam; and d Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

Objective: To evaluate whether ovarian reserve tests (ORTs) add prognostic value to patient characteristics, such as female age, in the prediction of excessive response to ovarian hyperstimulation in patients undergoing IVF, and whether their performance differs across clinical subgroups.

Design: Authors of studies reporting on basal FSH, antimüllerian hormone (AMH), or antral follicle count (AFC) in relation to ovarian response to ovarian hyperstimulation were invited to share original data. Random intercept logistic regression models were used to estimate added value of ORTs on patient characteristics, while accounting for between-study heterogeneity. Receiver operating characteristic regression analyses were performed to study the effect of patient characteristics on ORT accuracy.

Setting: In vitro fertilization clinics.

Patient(s): A total of 4,786 women for the main analysis, with a subgroup of 1,023 women with information on all three ORTs.

Intervention(s): None.

Main Outcome Measure(s): Excessive response prediction.

Result(s): We included 57 studies reporting on 32 databases. Female age had an area under the receiver operating characteristic curve of 0.61 for excessive response prediction. Antral follicle count and AMH significantly added prognostic value to this. A model with female age, AFC, and AMH had an area under the receiver operating characteristic curve of 0.85. The combination of AMH and AFC, without age, had similar accuracy. Subgroup analysis indicated that FSH performed significantly worse in predicting excessive response in higher age groups, AFC did significantly better, and AMH performed the same.
In women undergoing IVF, the development of a large number of oocytes occurs in one third of IVF cycles (1, 2). Such an excessive response may lead to poorer-quality embryos, lower chances of pregnancy, or cycle cancellation (3–9). Additionally, patients with an excessive response are at risk of developing ovarian hyperstimulation syndrome (OHSS), a potentially life threatening condition (10–12). To maximize safety and efficacy of assisted reproductive technology programs, there is a need to identify patients at risk of an excessive response at the start of IVF/intracytoplasmic sperm injection and to apply effective measures to prevent such an excessive response from occurring.

Several patient characteristics, such as a lean habitus, young age, and the presence of polycystic ovary syndrome (PCOS), have been identified as conditions that predispose patients to OHSS (13). Unfortunately, precise expressions of the predictive accuracy of these characteristics are not available. In contrast, ovarian reserve tests (ORTs), such as antimüllerian hormone (AMH), antral follicle count (AFC), and FSH, have been assessed for their value in the prediction of an excessive response (4, 6, 14–27). It is not clear, however, whether these ORTs add to predictive and readily available patient characteristics, of which female age is the most important.

Because ovarian reserve decreases with increasing age, it is conceivable that the predictive value of the ORTs is mutually dependent on female age. Alternatively, the accuracy of AFC may be different in women with a higher body mass index (BMI). Moreover, BMI could further influence the predictive accuracy by possibly reducing the biologic availability of recombinant FSH for ovarian stimulation, thereby creating spuriously reduced ovarian responses (28). Most predictive accuracy studies, however, had a limited sample size, lacking the power to evaluate patient characteristics as modifiers of accuracy in specific subgroups and the ability to analyze the added value of the ORTs on patient characteristics.

To overcome the problem of small studies with restricted power, the present study applied an individual patient data (IPD) meta-analysis approach. By aggregating data on the level of the individual patient, more precise estimates of accuracy, evaluations of added accuracy, and identification of accuracy modifiers become possible, while taking between-study heterogeneity into account appropriately.

**MATERIALS AND METHODS**

**Data Acquisition**

We searched the existing literature for studies on the value of FSH, AFC, and AMH in predicting IVF outcome. We expanded searches from conventional systematic reviews on the subject and another IPD meta-analysis (IPD-Individual patient data Meta-analysis on Poor response prediction with Ovarian Reserve Tests [IMPORT]) on poor response prediction; searches were updated to include studies up to the end of 2009 (14, 29–32).

Key words used in the systematic MEDLINE search included synonyms for in vitro fertilization (IVF, controlled ovarian stimulation, in vitro fertilisation) and synonyms for the various tests (FSH, follicle stimulating hormone, AFC, antral follicle count or number, AMH, antimüllerian hormone, müllerian inhibiting substance). Studies presenting data on ovarian response to hyperstimulation, at least one ORT, and at least one patient characteristic were eligible for the present review. All titles and abstracts were evaluated for eligibility by two authors (M.D. and S.L.B. or S.L.B. and J.v.D.). If necessary, the opinion of a third author was decisive (F.J.M.B.).

All authors of potentially eligible primary studies were informed about this IPD meta-analysis initiative 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 on the format of the data requested.

After data acquisition, all data were scrutinized on quality and consistency and, whenever possible, converted into a single format. Any issues or inconsistencies were checked with the original author. For a more detailed description of the IPD meta-analysis methodology the reader is referred to previous articles (33, 34).

Within all eligible studies, a comparison was made between those studies that could and those that could not be included. Sensitivity–specificity pairs for excessive response prediction were calculated for the ORTs under study, using the thresholds for excessive response that had been set in each study. Spearman correlations were then calculated for sensitivity–specificity pairs across studies, to ascertain that the differences in sensitivity and specificity levels between included and not-included studies were likely the result of different threshold levels used, thereby reducing the likelihood of bias in the final analysis.

All original studies either had approval of their local research ethics committee or were exempt from obtaining such approval.
owing to the nature of the study. We evaluated the quality of the included studies using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) checklist, supplemented by a number of items to evaluate the risk of bias in prognostic studies. Whenever a particular variable was missing in an individual database or in an individual case within a database, data were not imputed. Baseline characteristics were analyzed in the total IPD dataset and for each of the individual studies.

Definitions

An excessive response was defined as the retrieval of more than 15 oocytes. This cutoff was selected because the definition for excessive responsive in most primary studies varied between more than 14 and more than 16 oocytes (6, 17, 19, 23, 35–38). Furthermore, it has been shown that clinical pregnancy rates decline with the retrieval of more than 15 oocytes, thus arguing that it is an unfavorable condition (39). Duration of subfertility was defined as the period from cessation of oral contraceptives and/or start of unprotected intercourse until the first IVF attempt. In the included studies, patients had been stimulated according to local protocol, resulting in a wide range of daily FSH dosages. In almost all studies a starting dosage of at least 150 IU was given. This dosage is considered the optimal daily dosage in expected normal responders; with this dose it may be assumed that all patients received adequate stimulation, creating growth of all follicles sensitive to FSH within the time frame of exposure (40).

Predictive accuracy was defined as the ability of the model to distinguish excessive responders from cases with a normal or poor response. We calculated areas under the receiver operating characteristic (ROC) curve (AUCs) for normal or poor response. We calculated areas under the model to distinguish excessive responders from cases with a normal or poor response. We calculated areas under the ROC curve (AUCs) for normal or poor response. We calculated areas under the ROC curve (AUCs) for normal or poor response. We calculated areas under the ROC curve (AUCs) for normal or poor response. We calculated areas under the ROC curve (AUCs) for normal or poor response. We calculated areas under the ROC curve (AUCs) for normal or poor response. We calculated areas under the ROC curve (AUCs) for normal or poor response. We calculated areas under the ROC curve (AUCs) for normal or poor response.

Statistical Analysis

Analyses were done in two steps. First, the added value of ORTs on top of the patient characteristics age, BMI, and duration of subfertility was assessed. As a part of this analysis, we assessed whether these results may have been influenced by differences in study characteristics or daily FSH dosage administered. Second, we examined whether the predictive performance depends on the patient characteristics age, BMI, and duration of subfertility.

Prediction of an excessive response using ORTs and patient characteristics. To study whether ORTs have an added value on top of patient characteristics in the prediction of an excessive response, we used random intercept logistic regression models. The random intercept model takes heterogeneity into account by assuming that included studies share a common ROC for each ORT 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 AUCs of the models described above and evaluate the statistical significance of any differences.

Influence of age, BMI, and duration of subfertility on the accuracy of ORTs in excessive response prediction. To study whether the accuracy of ORTs in the prediction of excessive response is modified by patient age, BMI, or duration of subfertility, we used the ROC regression model proposed by Pepe, Janes, and colleagues (41, 42). This model allows us to study the effects of patient or disease characteristics on the classification accuracy of tests. In this model, the ORT ROC curves are modeled as a function of the covariates age, BMI, and duration of subfertility.
We assumed the effect of the covariate in this meta-analysis to be identical across studies, but, as in the previous analysis, the positivity threshold corresponding to each sensitivity–specificity pair was allowed to vary between studies, thereby correcting for any heterogeneity between studies. The areas under the corresponding ROC curves were calculated, to express the discriminatory capacity (accuracy) of the ORT in women in the respective subgroups.

Data were analyzed using SPSS 17.0 (SPSS Inc.) and R version 2.9.0 (http://www.r-project.org/). Random intercept logistic regression prediction models were created with the ‘Lme4’ library, using the Laplace approximation to the likelihood.

RESULTS

Data Acquisition

The MEDLINE search up to the end of 2009 delivered 2,551 hits, of which 125 were eligible for inclusion. In 22 studies the authors were untraceable, 33 authors did not reply after repeated effort, in 12 studies the data were lost, and 2 studies were not suited for the current analysis. This resulted in a total of 32 databases, used for the preparation of 57 or more manuscripts, which could be included in this IPD study. Twenty-seven had been previously included in the IPD–IMPORT study (32). Ten additional studies were identified from the systematic MEDLINE search. We invited these 10 extra authors and asked them, as well as the previous 27 studies, for permission to use their databases in the present analysis on excessive response prediction. Only four of these authors sent their data (11, 12, 25, 37), one of them submitting two separate databases (25). In total, 32 datasets could be included in the EXPORT study project database, with data from 5,251 study participants (Supplemental Table 1).

We were able to replicate the primary findings of the original study in 13 databases. In 12 cases, the study database we received contained a number of patients that differed from the publication, whereas in seven other databases there were slight inconsistencies with the baseline data as previously published. These inconsistencies were discussed with the corresponding author and could be resolved in all cases. Through this process, the level of consistency between the individual data and the data reported in the published articles was regarded sufficient for all included studies.

For the comparison of the four included and the six not-included studies, we attempted to calculate sensitivity and specificity of the ORTs in the prediction of excessive response. However, of the nonincluded studies only one reported sensitivity and specificity values for AFC in the prediction of an excessive response (23). Therefore, Spearman correlation could not be calculated. Nonetheless, for the majority of the studies this was performed in the IMPORT study (32), a related IPD study from the same research group focused on poor response prediction. In that study it was demonstrated that there was no difference in the correlations between sensitivity and specificity for included and not-included studies on poor response. Because there was no difference in poor response prediction, it is reasonable to assume that there is also no difference for excessive response prediction. We therefore assumed that no obvious bias has occurred for the present analysis by excluding studies on the basis of the availability of primary data. Baseline characteristics of the original studies are summarized in Supplemental Figure 3.

Data from 4,786 of the 5,251 women were suitable for the analysis of prediction of excessive response, of which 894 (19%) had an excessive response. In the other 465 women information on oocyte yield was missing. Baseline characteristics of the total study group are summarized in Table 1. The AUCs of the original studies for excessive response prediction are summarized in Supplemental Table 1.

Statistical Analyses

Prediction of an excessive response using ORTs and patient characteristics. For the model building exercises, we could use data of 1,023 women from 10 datasets for excessive response analysis. This was the number of women for whom all five variables of interest were known: age, AFC, AMH, FSH, and the number of oocytes retrieved after stimulation. Of the evaluated patient characteristics, age was the strongest single predictor of excessive response (odds ratio [OR] 0.89, 95% confidence interval [CI] 0.85–0.93), as shown in Table 2. Body mass index and duration of subfertility were not significantly predictive of excessive response (Supplemental Table 2).

We compared the ORTs using the random intercept logistic regression model in predicting excessive response. The ROC regression analysis showed a high accuracy for

### Table 1: Baseline characteristics from pooled data.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Total population (n = 4,786)</th>
<th>Excessive responders (n = 894)</th>
<th>Nonexcessive responder (n = 3,892)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female age (y)</td>
<td>34.4 (26.0–42.0)</td>
<td>32.5 (25.0–39.9)</td>
<td>34.7 (26.0–42.0)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>FSH (IU/L)</td>
<td>7.7 (3.8–14.0)</td>
<td>6.4 (3.5–10.1)</td>
<td>8.7 (3.9–16.0)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>AFC (n)</td>
<td>12.1 (3.0–25.6)</td>
<td>17.1 (6.0–32.0)</td>
<td>11.0 (3.0–22.0)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>AMH (ng/mL)</td>
<td>2.5 (0.1–7.6)</td>
<td>4.8 (1.3–10.2)</td>
<td>2.0 (0.1–5.7)</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>23.6 (18.6–30.1)</td>
<td>23.4 (18.5–29.4)</td>
<td>23.4 (18.6–30.1)</td>
<td>943</td>
</tr>
<tr>
<td>Duration of subfertility (y)</td>
<td>4.3 (1.3–10.0)</td>
<td>4.3 (1.5–10.0)</td>
<td>4.3 (1.2–10.0)</td>
<td>937</td>
</tr>
</tbody>
</table>

Note: Excessive response definition: >15 oocytes retrieved. Duration of subfertility: the period from the cessation of contraceptive methods or start of unprotected intercourse until the first IVF attempt. Values are presented as mean (5th percentile–95th percentile).

AMH (AUC 0.81, 95% CI 0.76–0.87) and for AFC (AUC 0.79, 95% CI 0.74–0.84), but only a moderate accuracy for FSH (AUC 0.66, 95% CI 0.60–0.73) (Fig. 1A).

The multivariable analyses demonstrated that a model including age, AFC, and AMH (AUC 0.85) had a significantly higher predictive accuracy than a model based on age alone (AUC 0.61; P < .001). Addition of FSH to this model did not further improve predictive accuracy (AUC 0.85; P = .73) (Fig. 1A). Interestingly, a single AMH or AFC test had a comparable accuracy (AUC 0.81 and 0.79, respectively). Addition of AMH to AFC and of AFC to AMH significantly improved accuracy (P < .001 or P = .003, respectively). A model combining these two tests resulted in an AUC of 0.85. Age did not add value to this model (P = .98). The ROC curves corresponding to the multivariable models are shown in Figure 1B.

**Effect of daily FSH dosage and study protocol on excessive response outcome.** Patients had been stimulated with a wide range of daily FSH dosages, according to their center’s local protocol. The mean daily FSH dosage was 204.28 IU (interquartile range 150–225 IU). Twenty-one women received daily FSH dosages <150 IU because of an expected excessive response (5 women received 75 IU, 14 women received 112.5 IU, and 2 women received 125 IU of daily FSH). Women who developed an excessive response tended to have received a lower starting dosage of FSH than women who did not develop an excessive response. The mean dosage was 201.75 IU in those women who developed an excessive response, vs. a mean dosage of 224.79 IU for women who did not have an excessive response (P value for difference < .001). Daily FSH dosage had a significant, negative association with excessive response development. A higher daily FSH dosage was associated with a lower chance of an excessive response in both the three-test study group and in the group as a whole (OR 0.99, P < .001). In the individual studies it was often not stated whether daily FSH dosage protocols were altered according to the results of the ORTs that were measured. Because it is very likely that this occurred and because it is further likely that different physicians acted differently to ORT results, adjusting at the level of the individual study was deemed to not be enough, and correction on the individual level was necessary. When daily FSH dosage was included in the multivariable model as an additional covariate (in addition to age and the ORTs), the ORs for age and the ORTs, adjusted for FSH dosage, remained basically unchanged. In the multivariable model for age, FSH, and daily FSH dosage, FSH had an OR of 0.86 (95% CI 0.83–0.90), the OR for AFC in the multivariable model for age, AFC, and daily FSH dosage remained 1.13 (95% CI 1.11–1.15), and the OR for AMH in the multivariable model for age, AMH, and FSH dosage was 1.55 (95% CI 1.45–1.66).

Study quality characteristics as scored by QUADAS checklist and supplemental questions are shown in Supplemental Figure 3. Overall, data were of high quality, with the exception of verification bias. This implies that the test results may have been known to the clinician taking decisions on patient management. None of the study characteristics that were assessed were associated with excessive response development (P value range .34–.89). Similarly, the ORs for age and the ORTs, adjusted for study characteristics, remained basically unchanged.

**Influence of age, BMI, and duration of subfertility on the accuracy of ORTs in excessive response prediction.** The results of the ROC regression model that studied the effect of several patient characteristics on the ROC curve of the ORTs in the prediction of an excessive response are shown in Table 3. The accuracy of FSH was significantly lower in women with a higher age (P = .01).

For a 20 year old the AUC for FSH was 0.66. In contrast, the AUC for a 30 year old was 0.59 and for a 40 year old was
The accuracy of AFC was significantly higher in women with a higher age ($P = .01$). For a 20-year-old woman the AUC for AFC was 0.64, for a 30 year old it was 0.71, and for a 40 year old it was 0.81. The discriminatory capacity of AMH in response prediction was not significantly influenced by age. Body mass index and duration of subfertility categories had no significant effect on the ROC curves, for any of the ORTs.

**DISCUSSION**

The results of the present IPD meta-analysis, with data from 32 individual studies, demonstrate that both AFC and AMH clearly add value to female age alone in the prediction of excessive response. Antimüllerian hormone and AFC in concert have high predictive accuracy, even without adding female age. The results also indicate that the performance of the ORTs may vary across patient subgroups, as determined by female age especially. At a higher female age FSH performs less well, whereas AFC performs better in older age groups. Because FSH performs the least well in excessive response prediction, this finding is not very relevant. For AFC the change in predictive accuracy with increasing age is more notable and results in an increased predictive accuracy, in terms of an increase in the AUC of approximately 0.26. However, this increase is only seen with big increments of female age (from 20 to 30 years or 30 to 40 years). With smaller increases in female age, such as between 31, 34, and 37 years (the 25th, 50th, and 75th percentiles of age and thus the most clinically relevant group) the increase in AUC is much smaller and less clinically relevant. In addition, the gain
in predictive accuracy is evenly spread over the entirety of the curve, thus limiting the margin of additive clinical value.

The results of this IPD meta-analysis are mostly in line with those from a previous, conventional systematic review and meta-analysis of ovarian reserve tests and excessive response (35) and another recent study in which AMH was able to accurately identify 79% of excessive responders (57). Our IPD approach allowed us to evaluate the added value of ORTs on top of female age and, moreover, allowed for the analysis of accuracy in subgroups of women defined by age, BMI, or duration of subfertility. Although ORT adds value to female age in predicting excessive response, age adds little to nothing to the accuracy of the prediction based on the ORTs. It does, however, does seem to influence the accuracy of some ORTs.

The results of this IPD meta-analysis also suggest that age influences the accuracy of AFC and basal FSH. Although ovarian reserve decreases with age, the AFC is believed to reflect the true level of the quantitative ovarian reserve directly, in contrast to basal FSH, which constitutes an indirect marker of follicle numbers. Indeed, in older women the prevalence of excessive response may become too low for any test to gain sufficient accuracy, and this may be especially true for FSH. For AFC the change in accuracy may be significant only from the statistical point of view, without actual implications for clinical practice, and without an obvious explanatory mechanism.

A challenge with the IPD approach is collecting sufficient data. For the present study databases of 60 of the eligible 125 manuscripts were obtained. We were unable to reach a number of authors, primarily because of inaccurate contact information or because authors did not reply to the e-mail addresses provided. Older data were often lost or in a format that could no longer be read. Studies to investigate the possibility of combining IPD data with aggregated data are ongoing (58). To compare included and excluded studies, we aimed to calculate Spearman correlation coefficients for the included and nonincluded studies. Unfortunately, of the nonincluded studies only one reported sensitivity and specificity values for AFC in the prediction of an excessive response. Therefore, Spearman correlation could not be calculated. However, for 27 of 32 studies a Spearman correlation was calculated from a previous IPD meta-analysis on poor response prediction, and this showed that there was no difference (14). Because there is no difference in poor response prediction, it is reasonable to assume that there is also no difference for excessive response prediction. Therefore, we believe that the current number of participants and amount of data allowed us to analyze a valid selection of all the available data. It would have been interesting to add PCOS as a candidate predictor in our uni- and multivariate analyses because women with PCOS have been found to be prone to establishing OHSS after IVF treatment (13). However, in the majority of studies, PCOS was one of the exclusion criteria, and from those studies that included and recorded PCOS a mere 131 women had PCOS.

Although the present IPD meta-analysis included studies up to the end of 2009, the results of more recent studies on the value of ORTs in predicting ovarian response are still in agreement with our findings of the present IPD meta-analysis. Two recent studies in an IVF setting (57, 59) and three studies performed in oocyte donors or breast cancer patients undergoing oocyte cryopreservation all show an AUC of approximately 0.80 for AMH in excessive response prediction (60–62).

Using original data of a number of studies comes with between-study heterogeneity. The incorporation of ovarian reserve tests and restrictions based on test results in everyday IVF practice has led to selection bias in some study populations. Heterogeneity found in the included studies pertained to differences in IVF indications, access to IVF resources, differing treatment protocols, variability in embryo laws, and discordant definitions of ongoing pregnancy. There is also a variation in hormone assays and AFC sizes measured, for which no international consensus exists to correct for these differences. Consequently, no cutoff values for these tests could be used or mentioned. The most valuable method of obtaining such cutoff values for clinical practice is through randomized controlled trials, which are underway at the moment (63). We have used random intercept logistic regression as well as the ROC regression model by Janes, Pepe, and colleagues (41, 42), in which pertinent heterogeneity between studies is accounted for.

The clinical value of excessive response prediction will depend on the consequences for clinical management. Several studies have looked at the effect of individualized treatment protocols. By providing women with personally tailor-made stimulation protocols (i.e., with a lower daily FSH dosage), it is attempted to keep the oocyte yield between 5 and 12 oocytes. At present the evidence is inconclusive upon the effectiveness of such personalized treatment regimens based on a priori prediction of ovarian reserve (50, 51). In the study of Popovic-Todorovic et al. (51) the use of an individualized protocol resulted in a larger number of normal responders but a similar number of excessive responders. In contrast, Olivennes et al. (64) demonstrated that lower individualized dosage protocols allow for a similar oocyte yield, implantation rate, and pregnancy compared with higher dosage protocols. A third study (65) showed no difference in the number of mature oocytes retrieved or in the occurrence of OHSS between patients that were randomly assigned to receive 225 IU or 300 IU of FSH. Last, it has been suggested that individualization of stimulation protocols dose on the basis of ovarian reserve tests is expected to be cost effective in IVF populations (66).

On the basis of the present study we cannot speculate about associations between FSH dosage and excessive response prevention. A significant association between daily FSH dosage and excessive response was found, with women with lower daily FSH dosages having higher chances of excessive response. This association reflects physician behavior, whereby lower daily FSH dosages are preemptively prescribed according to specific patient characteristics, ORT results, or any comorbidity in anticipation of an excessive response. This suggests a form of selection bias, whereby the accuracy of ORTs or patient characteristics in the prediction of an excessive response is actually higher than currently reported, because some excessive responses may have been prevented by prescribing lower daily FSH dosages. The high response despite
a low daily FSH dosage can be explained by the presence of a large number of follicles, with a sensitivity for FSH close to the FSH threshold (67). More prospectively collected evidence, in the form of large-scale randomized control trials, is needed to demonstrate whether an individualized treatment protocol based on ORTs and patient characteristics is a truly effective strategy in the prevention of an excessive response; a protocol for such a randomized control trial was recently published (63).

In conclusion, this IPD meta-analysis shows that AFC and AMH add predictive accuracy to age in the prediction of an excessive response. A model combining these ORTs provides good predictive accuracy, without the necessity of including female age. The performance of FSH and AFC, but not AMH, was influenced by female age but not by BMI or duration of subfertility. However, the performance across subgroups with small increments in female age seemed not to be sufficiently altered to be recognized as clinically relevant. The high predictive accuracy for both AMH and AFC or a combination of both urges the need for studies that examine the effect of ORT-based dose adaptations in which efficacy of treatment, costs, and response normalization is analyzed.


REFERENCES

4. Luna M, Grunfeld L, Mukherjee T, Sandler B, Copperman AB. Moderately elevated levels of basal follicle-stimulating hormone in young patients predict low ovarian response, but should not be used to disqualify patients from attempting in vitro fertilization. Fertil Steril 2007;87:782–7.
21. Liu KE, Greenblatt EM. Elevated day 3 follicle-stimulating hormone/luteinizing hormone ratio >0.2 is associated with higher rates of cancellation in in vitro fertilization-embryo transfer cycles. Fertil Steril 2008;90:297–301.


SUPPLEMENTAL FIGURE 1

Flowchart of included studies.

SUPPLEMENTAL FIGURE 2

A Number of patients per study

B Incidence of an excessive response per study

SUPPLEMENTAL FIGURE 2 Continued

LEGEND:
1 Kwee
2 Ng 2000
3 Ng 2005
4 Caroppo
5 Anderson
6 Klikkert
7 Nelson
8 Marce
9 Bancal
10 Tomás
11 Greenblatt
12 Multikrishna 2004
13 Multikrishna 2005
14 Ashrafi
15 Erdem
16 McIvane
17 Popovic 2003a
18 Popovic 2003b
19 Vladirov
20 La Marca
21 van der Linden
22 Elder-Geva
23 Jayaprakasan
24 Smeeck 2007
25 Copperman
26 Ebner
27 van Rooij
28 Fesour
29 Afshoonian
30 Gnoth
31 Nardo "unpublished"
32 Nardo 2008


Baseline characteristics of the included studies. (A) Number of patients per study. (B) Prevalence of an excessive response per study. (C) Mean, 5th percentile, and 95th percentile of the patient characteristics female age, BMI, and duration of subfertility for each individual study. (D) Mean, 5th percentile, and 95th percentile of ovarian reserve tests FSH, AFC, and AMH for each individual study.

Study characteristics according to QUADAS. Characteristics of all included studies evaluated with the QUADAS checklist. Note that QUADAS was set up for diagnostic studies, and these are all prognostic studies. Therefore, questions regarding reference test could not be answered. Some questions specific for ovarian reserve testing and fertility studies were added. All studies were cohort studies, with the majority prospectively set up. All studies analyzed the results per cycle, and some studies analyzed more cycles per couple, in which case only the first cycle was analyzed.

<table>
<thead>
<tr>
<th>Study</th>
<th>FSH AUC (95% CI)</th>
<th>FSH N</th>
<th>Assay</th>
<th>AFC AUC (95% CI)</th>
<th>AFC N</th>
<th>Criteria (mm)</th>
<th>AMH AUC (95% CI)</th>
<th>AMH N</th>
<th>Assay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alfatoonian</td>
<td>0.60 (0.50-0.69)</td>
<td>143</td>
<td>12</td>
<td>0.96 (0.93-0.99)</td>
<td>143</td>
<td>2-6</td>
<td>0.94 (0.90-0.98)</td>
<td>143</td>
<td>DSL</td>
</tr>
<tr>
<td>Anderson</td>
<td>0.92 (0.99-1.00)</td>
<td>46</td>
<td>11</td>
<td>0.61 (0.67-0.85)</td>
<td>46</td>
<td>2-10</td>
<td>NA</td>
<td>50</td>
<td>NA</td>
</tr>
<tr>
<td>Ashrafi</td>
<td>0.59 (0.31-0.87)</td>
<td>50</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Banci</td>
<td>0.61 (0.54-0.68)</td>
<td>505</td>
<td>6</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Caroppi</td>
<td>0.81 (0.72-0.90)</td>
<td>76</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Copperman</td>
<td>0.65 (0.60-0.69)</td>
<td>570</td>
<td>5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Ebner</td>
<td>0.61 (0.46-0.75)</td>
<td>127</td>
<td>NA</td>
<td>NA</td>
<td>0.82 (0.74-0.90)</td>
<td>135</td>
<td>BC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eldar-Geva</td>
<td>0.71 (0.57-0.85)</td>
<td>52</td>
<td>5</td>
<td>0.88 (0.75-1.00)</td>
<td>36</td>
<td>2-10</td>
<td>0.75 (0.62-0.88)</td>
<td>54</td>
<td>BC</td>
</tr>
<tr>
<td>Erdem</td>
<td>0.77 (0.57-0.97)</td>
<td>24</td>
<td>5</td>
<td>0.85 (0.70-1.00)</td>
<td>24</td>
<td>2-8</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Freour</td>
<td>0.58 (0.41-0.73)</td>
<td>62</td>
<td>NA</td>
<td>NA</td>
<td>0.70 (0.55-0.86)</td>
<td>64</td>
<td>BC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gnoth</td>
<td>0.64 (0.51-0.78)</td>
<td>122</td>
<td>NA</td>
<td>NA</td>
<td>0.87 (0.79-0.95)</td>
<td>134</td>
<td>DSL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greenblatt</td>
<td>0.67 (0.59-0.74)</td>
<td>261</td>
<td>5</td>
<td>0.69 (0.61-0.77)</td>
<td>223</td>
<td>2-8</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Jayaprakasan</td>
<td>0.74 (0.57-0.91)</td>
<td>100</td>
<td>NA</td>
<td>0.82 (0.70-0.95)</td>
<td>100</td>
<td>2-10</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Klinkert</td>
<td>0.42 (0.30-0.55)</td>
<td>212</td>
<td>4</td>
<td>0.45 (0.33-0.57)</td>
<td>221</td>
<td>2-5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Kwee</td>
<td>0.79 (0.70-0.88)</td>
<td>109</td>
<td>1</td>
<td>0.87 (0.82-0.96)</td>
<td>109</td>
<td>2-10</td>
<td>0.84 (0.76-0.92)</td>
<td>105</td>
<td>DSL</td>
</tr>
<tr>
<td>La Marca</td>
<td>NA</td>
<td>71</td>
<td>8</td>
<td>No &gt;15</td>
<td>71</td>
<td>2-10</td>
<td>0.90 (0.76-1.00)</td>
<td>118</td>
<td>BC</td>
</tr>
<tr>
<td>Mclveen</td>
<td>No &gt;15</td>
<td>71</td>
<td>8</td>
<td>No &gt;15</td>
<td>71</td>
<td>2-10</td>
<td>0.90 (0.76-1.00)</td>
<td>118</td>
<td>BC</td>
</tr>
<tr>
<td>Mercia</td>
<td>NA</td>
<td>62</td>
<td>0.62 (0.42-0.83)</td>
<td>65</td>
<td>2-5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Muttukrishna 1</td>
<td>0.81 (0.59-1.00)</td>
<td>66</td>
<td>7</td>
<td>NA</td>
<td>0.92 (0.83-1.00)</td>
<td>66</td>
<td>BC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muttukrishna 2</td>
<td>0.67 (0.52-0.82)</td>
<td>68</td>
<td>7</td>
<td>0.84 (0.73-0.94)</td>
<td>68</td>
<td>NA</td>
<td>0.73 (0.56-0.91)</td>
<td>68</td>
<td>BC</td>
</tr>
<tr>
<td>Nardo 1</td>
<td>0.65 (0.53-0.77)</td>
<td>135</td>
<td>5</td>
<td>0.71 (0.59-0.83)</td>
<td>123</td>
<td>2-5</td>
<td>0.74 (0.64-0.83)</td>
<td>135</td>
<td>DSL</td>
</tr>
<tr>
<td>Nardo 2</td>
<td>0.68 (0.59-0.77)</td>
<td>145</td>
<td>13</td>
<td>0.71 (0.63-0.80)</td>
<td>145</td>
<td>2-5</td>
<td>0.79 (0.72-0.87)</td>
<td>145</td>
<td>DSL</td>
</tr>
<tr>
<td>Nelson</td>
<td>0.64 (0.58-0.71)</td>
<td>338</td>
<td>5</td>
<td>NA</td>
<td>0.88 (0.82-0.91)</td>
<td>319</td>
<td>DSL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ng 1</td>
<td>0.70 (0.56-0.83)</td>
<td>131</td>
<td>2</td>
<td>0.80 (0.70-0.90)</td>
<td>131</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Ng 2</td>
<td>0.72 (0.56-0.83)</td>
<td>109</td>
<td>5</td>
<td>0.77 (0.68-0.85)</td>
<td>127</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Popovic 1</td>
<td>0.62 (0.54-0.71)</td>
<td>256</td>
<td>1</td>
<td>0.71 (0.63-0.80)</td>
<td>256</td>
<td>2-5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Popovic 2</td>
<td>0.62 (0.50-0.73)</td>
<td>143</td>
<td>1</td>
<td>0.76 (0.67-0.86)</td>
<td>143</td>
<td>2-5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Smeenk 1</td>
<td>0.54 (0.40-0.68)</td>
<td>80</td>
<td>10</td>
<td>0.66 (0.530.79)</td>
<td>80</td>
<td>2-10</td>
<td>0.71 (0.57-0.84)</td>
<td>80</td>
<td>BC</td>
</tr>
<tr>
<td>Tomas</td>
<td>NA</td>
<td>82</td>
<td>0.82 (0.72-0.91)</td>
<td>160</td>
<td>2-5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>Van Rooij</td>
<td>0.68 (0.58-0.79)</td>
<td>215</td>
<td>10</td>
<td>0.86 (0.79-0.93)</td>
<td>215</td>
<td>2-5</td>
<td>0.87 (0.77-0.97)</td>
<td>215</td>
<td>BC</td>
</tr>
<tr>
<td>Van der Linden</td>
<td>0.82 (0.72-0.92)</td>
<td>124</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Vladimirov 2</td>
<td>0.67 (0.48-0.87)</td>
<td>39</td>
<td>9</td>
<td>0.74 (0.52-0.97)</td>
<td>39</td>
<td>2-10</td>
<td>0.80 (0.67-0.93)</td>
<td>39</td>
<td>BC</td>
</tr>
</tbody>
</table>

Note: FSH assays: 1 = Immunometric, Delia; 2 = automated chemiluminescence, ACS180, Bayer; 3 = immunoradiometric, Immunotech; 4 = immunoometric assay, Chiron Diagnostics; 5 = Immulite semiautomated, DPC; 6 = Enzymun-FSH test, Boehringer Mannheim; 7 = immunoradiometric assay, DPC; 8 = chemiluminescence detection, Adiva Centaur, Bayer; 9 = electrochemiluminescence immunoassay, Roche Elecsys; 10 = fluorescence immunoenzymometric, AsYLM, Abbott; 11 = double antibody assay, Organon; 12 = IDCS, Korbach; 13 = Roche E170 automated immunoassay, AMH assay: DSL = Diagnostic Systems Laboratories; BC = Beckman Coulter; NA = not available. Areas under the curve (AUC) from original studies for prediction models of ovarian reserve tests for the prediction of an excessive response.

**SUPPLEMENTAL TABLE 2**

Univariable and multivariable models of patient characteristics in the prediction of an excessive response.

<table>
<thead>
<tr>
<th>Model</th>
<th>Three-test study group (n = 1,023)</th>
<th></th>
<th>Total study group (n = 4,786)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td>95% CI</td>
<td>P value</td>
<td>OR</td>
</tr>
<tr>
<td>Univariable model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.89</td>
<td>0.85–0.93</td>
<td>&lt;.001</td>
<td>0.90</td>
</tr>
<tr>
<td>BMI</td>
<td>0.98</td>
<td>0.93–1.03</td>
<td>.405</td>
<td>1.00</td>
</tr>
<tr>
<td>Duration</td>
<td>0.98</td>
<td>0.90–1.06</td>
<td>.555</td>
<td>0.97</td>
</tr>
<tr>
<td>Multivariable model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age and BMI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.91</td>
<td>0.87–0.95</td>
<td>&lt;.001</td>
<td>0.9</td>
</tr>
<tr>
<td>BMI</td>
<td>0.99</td>
<td>0.93–1.04</td>
<td>.616</td>
<td>1.00</td>
</tr>
<tr>
<td>Age and duration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.90</td>
<td>0.85–0.94</td>
<td>&lt;.001</td>
<td>0.89</td>
</tr>
<tr>
<td>Duration</td>
<td>1.01</td>
<td>0.93–1.10</td>
<td>.750</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: Duration = duration of subfertility.