Prediction of an excessive response from patient characteristics and ovarian reserve tests and comparison in subgroups: an Individual Patient Data Meta-Analysis

Running title: Predicting excessive response to IVF with ORTs

The EXPORT* study group

S.L. Broer[§] and M. Dólleman[§]**, J. van Disseldorp, K.A. Broeze, B.C. Opmeer, A. Aflatoonian, R.A. Anderson, M. Ashrafi, L. Bancsi, E. Caroppo, A.B. Copperman, T. Ebner, T. Eldar-Geva, M. Erdem, T. Freour, C. Gnoth, E.M. Greenblatt, K. Jayaprakasan, N. Raine-Fenning, E. Klinkert, J. Kwee, A. La Marca, C.B. Lambalk, M. McIlveen, L. Mohiyiddeen, L.T. Merce, S. Muttukrishna, L.G. Nardo, S.M. Nelson, H.Y. Ng, B. Popovic-Todorovic, J.M.J. Smeenk, C. Tomás, P.J.Q. Van der Linden, I.A. van Rooij, I.K. Vladimirov, P. Bossuyt, M.J.C. Eijkemans, B.W. Mol and F.J.M. Broekmans, on behalf of the IPD-EXPORT* study group

§ Both authors contributed equally

* Excessive Response Prediction using Ovarian Reserve Tests

** To whom correspondence should be addressed: University Medical Center Utrecht, room 05.126, P.O. Box 85500, 3508 GA, Utrecht, the Netherlands.
E-mail: M.Dolleman-2@umcutrecht.nl.

Table of contents

Capsule

This IPD meta-analysis demonstrates that AFC and AMH add value to age in predicting excessive response to ovarian hyperstimulation and that the accuracy of some ORTs is affected by age.

Abstract

Introduction: An excessive response to ovarian hyperstimulation during IVF is associated with patient discomfort and complications. This individual patient data (IPD) meta-analysis, evaluates whether ovarian reserve tests (ORTs) add prognostic value to patient characteristics, like female age in the identification of excessive responders, and whether their performance differs across clinical subgroups.

Methods: We searched for studies published until December 2009 of basal FSH, AMH or AFC in relation to ovarian response to ovarian hyperstimulation and authors were invited to share their original data. Random intercept logistic regression models were used to estimate the added value of the ORTs on patient characteristics, while accounting for between study heterogeneity. ROC regression analyses were performed to study the effect of specific patient characteristics on the accuracy of the ORTs.

Results: Thirty-two databases could be included (n=5,251). Age had an area under the ROC curve (AUC) of 0.61 for excessive response prediction. AFC and AMH significantly added prognostic value to age (P-value for each <0.001). A model with age, AFC and AMH had an AUC of 0.85. The combination AMH and AFC, without age had similar accuracy (P=0.98). The subgroup analysis showed that FSH performed worse (P=0.01) in predicting excessive response in higher age groups, AFC did better (P=0.01) and AMH performed about the same (p=0.14).

Conclusion: This IPD meta-analysis demonstrates that AFC and AMH add value to female age in the prediction of excessive response and that, for some ORTs, the discriminatory performance is affected by female age. ORTs, and specifically AMH, may thus be useful for excessive response prediction in IVF-populations.

Introduction

In women undergoing in vitro fertilization (IVF), the development of a large number of oocytes complicates up to thirty percent of IVF cycles (Delvigne and Rozenberg, 2002). Such an excessive response may lead to poorer quality embryos, lower chances of pregnancy, or cycle cancellation (Baart, Martini *et al.*, 2006;Heijnen, Eijkemans *et al.*, 2007) (Verberg, Eijkemans *et al.*, 2009) (van der Gaast, Eijkemans *et al.*, 2006). Additionally, patients with an excessive response are at risk of developing ovarian hyperstimulation syndrome (OHSS), a potentially life threatening condition (Fauser, Diedrich *et al.*, 2008) . To maximize safety and efficacy of assisted reproductive technology (ART) programs, there is a need to identify patients at risk of an excessive response at the start of IVF/ICSI treatment, 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 (Ho, Lee *et al.*, 2003). Unfortunately, precise expressions of the predictive accuracy of these characteristics are not available. In contrast, ovarian reserve tests (ORTs), such as Anti-Müllerian Hormone (AMH), Antral Follicle Count (AFC) and Follicle Stimulation Hormone (FSH) have been assessed for their value in the prediction of an excessive response (Broer, Mol *et al.*, 2010) (van Rooij, Broekmans *et al.*, 2002a) (Eldar-Geva, Ben Chetrit *et al.*, 2005b) (Nakhuda, Chu *et al.*, 2006) (Riggs, Duran *et al.*, 2008) (Nardo, Gelbaya *et al.*, 2009). It is not clear, however, what ORTs add to predictive and readily available patient characteristics, such as age.

As ovarian reserve decreases with age, it is conceivable that the predictive value of the ORTs also depends on female age. Alternatively, the accuracy of the antral follicle count may be more complicated in women with a higher BMI. Moreover, BMI could further influence the predictive accuracy by possibly reducing the biologic availability of recombinant FSH for ovarian stimulation, and thereby creating spuriously reduced ovarian responses (Steinkampf, Hammond *et al.*, 2003). 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 current study applied an individual patient database (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 becomes possible while taking between study heterogeneity into account appropriately.

Material and Methods

Data acquisition

We searched for 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-IMPORT) on poor response prediction; searches were updated to include studies up to the end of 2009. (Broekmans, Kwee *et al.*, 2006) (Broer, Mol *et al.*, 2009) (Broer, Eijkemans *et al.*, 2011) (Broer, 2011) (Broer, Mol *et al.*, 2010)**.**

Keywords 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, Anti-Müllerian Hormone, Müllerian inhibiting substance). Studies presenting data on ovarian response to hyperstimulation, at least one ovarian reserve test (ORT) and at least one patient characteristic were eligible for the current review. All titles and abstracts were evaluated for eligibility by two authors (MD and SB or SB and JvD). If necessary the opinion of a third author was decisive (FB).

All authors of potentially eligible primary studies were informed about this individual patient data (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 papers (Broeze, Opmeer *et al.*, 2009;Broeze, Opmeer *et al.*, 2011).

Within all eligible studies, a comparison was made between those studies that could and those that could not be included. Sensitivity and 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 and 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.

We evaluated the quality of the included studies using the QUADAS checklist, supplemented by a number of items to evaluate the risk of bias in prognostic studies. 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 cut-off was selected as the definition for excessive responsive in most primary studies varied between more than 14 and more than 16 oocytes (Broer, Dolleman *et al.*, 2011). 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 FSH dosages. In almost all studies a starting dosage of at least 150 International Units (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 (Sterrenburg, Veltman-Verhulst *et al.*, 2011)

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–Operator

Characteristic Curve (ROC-AUC) for the ORTs in the prediction of excessive response for each individual study and for the pooled studies were calculated as a summary statistic of predictive accuracy.

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 FSH dosage administered. Secondly, 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 are a random sample of a potential universe of studies, and that between-study variation in the incidence of excessive response in this universe can be described by a normal distribution on the log odds scale. These models were created to quantitatively estimate the added value that ORTs have on patient characteristics in predicting an excessive response. It provides both an estimate of the summary predictive effect as well as of the variance of this distribution.

Three different sets of models were used for the prediction of excessive response. The first set of models included the patient characteristics female age, BMI, and duration of subfertility. In the second set of models, the predictive capacity of each of the individual ovarian reserve tests (FSH, AFC and AMH) was estimated. In the third set of multivariate models, the added value of combinations of ovarian reserve tests on top of patient characteristics was evaluated.

The next step was to construct receiver operating characteristic (ROC) curves to express the predictive accuracy of each combination of predictive variables in distinguishing excessive responders

from the rest. With each of the random intercept logistic regression models, we calculated the probability of an excessive response. By moving the positivity threshold from 0 to 1, we could then calculate sensitivity-specificity pairs for each model. Based on these, we plotted stratified ROC curves with the ROC regression model as proposed by Janes and Pepe (Janes, Longton *et al.*, 2009;Pepe, Longton *et al.*, 2009). This model assumes that 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.

Because not all studies in this meta-analysis had included data for all three ORTs, we constructed prediction models using those databases from the total dataset that included the corresponding ovarian reserve tests (FSH, AFC and AMH) and age to allow for a direct comparison. The results of all analyses in the three-test study subgroup were verified in the total study group.

To account for between study differences in FSH dosage protocols and their potential effect on excessive response, we repeated the analyses as described above while adding starting FSH dosage as a covariate. In a similar fashion, we included study design features, as identified by the QUADAS checklist, as covariates in our models, in order to evaluate whether differences in FSH dosage or study design influenced the observed associations between ORT, patient characteristics and the outcome excessive response (Whiting, Rutjes *et al.*, 2011).

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 and Janes (Janes, Longton *et al.*, 2009;Pepe, Longton *et al.*, 2009). 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 (AUC) were calculated in order to express the discriminatory capacity (accuracy) of the ORT in women in the respective subgroups.

Data were analyzed using SPSS 17.0 (SPSS Inc., Chicago, Il, USA) 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

A total of 32 databases, used for the preparation of 57 or more manuscripts, could be included in this IPD-study. Twenty-seven had been previously included in the IPD-IMPORT study (Broer, 2011). Ten additional studies were identified from the systematic MEDLINE search. We invited these authors and asked them for permission to use their databases in the present analysis on excessive response prediction. Only four of these authors sent their data (Aflatoonian, Oskouian *et al.*, 2009) (Freour, Mirallie *et al.*, 2007) (Gnoth, Schuring *et al.*, 2008) (Nardo, Gelbaya *et al.*, 2009); one of them submitted two separate databases (Nardo, Gelbaya *et al.*, 2009). In total 32 datasets could be included in the EXPORT study project database, with data from 5,251 study participants (Figure 1).

With the original data 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 7 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 most cases. Through this process, the level of consistency between the individual data and the data reported in the published manuscripts was regarded sufficient for all included studies.

EXPORT studies

Aflatoonian *et al.,* 2009 ; (Ashrafi, Madani *et al.*, 2005;Yong, Baird *et al.*, 2003;Bancsi, Huijs *et al.*, 2000;Caroppo, Matteo *et al.*, 2006;Luna, Grunfeld *et al.*, 2007;Eldar-Geva, Ben Chetrit *et al.*, 2005a;Erdem, Erdem *et al.*, 2004;Liu and Greenblatt, 2008;Jayaprakasan, Hilwah *et al.*, 2007;Klinkert, Broekmans *et al.*, 2005;Kwee, Elting *et al.*, 2003;La Marca, Giulini *et al.*, 2007;McIlveen, Skull *et al.*, 2007;Merce, Barco *et al.*, 2007;Ng, Tang *et al.*, 2000;Ng, Chan *et al.*, 2005;Muttukrishna, Suharjono *et al.*, 2004;Muttukrishna, McGarrigle *et al.*, 2005;Nelson, Yates *et al.*, 2007;Popovic-Todorovic, Loft *et al.*, 2003a;Popovic-Todorovic, Loft *et al.*, 2003c;Smeenk, Stolwijk *et al.*, 2000;Smeenk, Sweep *et al.*, 2007;Tomas-C, Nuojua-Huttunen- *et al.*, 1997;van Swieten, Leeuw-Harmsen *et al.*, 2005;van Rooij, Broekmans *et al.*,

For the comparison of the included and not included studies, we attempted to calculate sensitivity and specificity of the ORTs in the prediction of excessive response. However, of the non-included studies only one reported sensitivity and specificity values for AFC in the prediction of an excessive response. Therefore, Spearman correlation could not be calculated. Nonetheless, for the majority of the studies this was performed in the IMPORT study (Broer, 2011), 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 non-included studies on poor response. Since 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 based on the

availability of primary data. Baseline characteristics of the original studies are summarized in Table A-1 of the online supplementary data.

Data from 4,786 out of the 5,251 women were suitable for the analysis of prediction of excessive response, of which 894 (19%) had an excessive response. The other women were not suitable as the primary outcome was ongoing pregnancy and not oocyte yield. Baseline characteristics of the total study group are summarized in 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 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 (OR 0.89; 95% CI: 0.85 to 0.93). BMI and duration of subfertility were not significantly predictive of excessive response (Addendum Table A-IV).

We compared the ORTs using the random intercept logistic regression model in predicting excessive response (see Table 2). The ROC regression analysis showed a high accuracy for AMH (AUC 0.81: 95% CI 0.76 to0.87) and for AFC (AUC 0.79: 95% CI 0.74 to 0.84), but only a moderate accuracy for FSH (AUC 0.66: 95% CI 0.60 to 0.73) (Table 3).

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=<0.001). Addition of FSH to this model did not further improve predictive accuracy (AUC 0.85; p = 0.73) (Table 3). 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 = <0.001 or p=0.003, respectively). A model combining these two tests resulted in an AUC of 0.85. Age did not add value to this model (p = 0.98). The ROC curves corresponding to the multivariable models are shown in Figure 2.

Effect of FSH dosage and study protocol on excessive response outcome

Patients had been stimulated with a wide range of FSH dosages according to their center's local protocol. The mean FSH dosage was 204.28 IU (IQR=150-225 IU). 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 versus a mean dosage of 224.79 IU for women who did not have an excessive response (p-value for difference <0.001). FSH dosage had a significant, negative association with excessive response development. A higher 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<0.001). When FSH dosage was included in the multivariable model as an additional covariate (in addition to age and the ORTs) the odds-ratios for age and the ORTs, adjusted for FSH dosage, remained basically unchanged.

Study quality characteristics as scored by QUADAS checklist and supplemental questions are shown in Figure 2. 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. Additional study characteristics with regard to sampling, data collection and study design are shown in Table A-I, addendum. None of the study characteristics that were assessed were associated with excessive response development (p-value range 0.34-0.89). Similarly, the odds-ratios 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 which 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 4. The accuracy of FSH was significantly lower in women with a higher age (p = 0.01).

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 0.52 for a 40 year old. The accuracy of AFC was significantly higher in women with a higher age (p =

0.01). 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. BMI 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. AMH 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, while AFC performs better in younger age groups. As 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 area under the curve, 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 25^{th}, 50^{th} and 75^{th} percentiles of age and thus the most clinically relevant group) the increase in AUC is much smaller. 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 (Broer, Dolleman *et al.*, 2011) and another recent study in which AMH was able to accurately identify 79% of excessive responders (Anckaert, Smitz *et al.*, 2012). 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 to age, BMI or duration of subfertility. While 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 current 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 (Riley, Dodd *et al.*, 2008). To compare included and excluded studies we aimed to calculate Spearman correlation coefficients for the included and non-included studies. Unfortunately, of the non-included 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 out 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, (Broer, Mol *et al.*, 2010). Since 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.

Although the current 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 this current IPD-meta-analysis. Two recent studies in an IVF setting (Anckaert, Smitz *et al.*, 2012) (Andersen, Witjes *et al.*, 2011) and three studies performed in oocyte donors or breast cancer patients undergoing oocyte cryopreservation all show an AUC of around 0.80 for AMH in excessive response prediction(Lee, Ozkavukcu *et al.*, 2011) (Nakhuda, Douglas *et al.*, 2011) (Riggs, Kimble *et al.*, 2011).

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 cut-off values for these tests could be used or mentioned. We have used random intercept logistic regression as well as the ROC regression model by Janes and Pepe et al. (Janes, Longton *et al.*, 2009;Pepe, Longton *et al.*, 2009) 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 FSH dosage, it is attempted to keep the oocyte yield between 5-12 oocytes. At present, the evidence is inconclusive upon the effectiveness of such personalized treatment regimens based on a priori prediction of ovarian response (Popovic-Todorovic, Loft *et al.*, 2003d;Popovic-Todorovic, Loft *et al.*, 2003b). In the study of Popovic-Todorovic the use of an individualized protocol resulted in a larger number of normal responders but a similar number of excessive responders (Popovic-Todorovic, Loft *et al.*, 2003b). In contrast, Olivennes et al. demonstrated that lower individualized dosage protocols allow for a similar oocyte yield, implantation rate and pregnancy compared to higher dosage protocols (Olivennes, Howies *et al.*, 2011). A third study 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 (Jayaprakasan, Hopkisson *et al.*, 2010).

Based on the current study we cannot speculate about associations between FSH dosage and excessive response prevention. A significant association between FSH dosage and excessive response was found, with women with lower FSH dosages having higher chances of excessive response. This association probably reflects physician behavior, where lower FSH dosages are preemptively prescribed guided by specific patient characteristics, ORT results, or any comorbidity in anticipation

of an excessive response. This suggests a form of selection bias, where the accuracy of ORTs or patient characteristics in the prediction of an excessive response is actually higher than currently reported, as some excessive responses may have been prevented by prescribing lower FSH dosages. The high response despite a low FSH dosage can be explained by the presence of a large number of follicles with a sensitivity for FSH close to the FSH threshold (Van der Meer, Hompes *et al.*, 1998). 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 an truly effective strategy in the prevention of an excessive response, a protocol for such a randomized control trial was recently published (van Tilborg, Eijkemans *et al.*, 2012).

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 to include 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.

Acknowledgements

List of contributions

Data collection: A. Aflatoonian, R.A. Anderson, M. Ashrafi, L. Bancsi, E. Caroppo, A.B. Copperman, T. Ebner, T. Eldar-Geva, M. Erdem, T. Freour, C. Gnoth, E.M. Greenblatt, K. Jayaprakasan, N. Raine-Fenning, E. Klinkert, J. Kwee, A. La Marca, C.B. Lambalk, M. McIlveen, L. Mohiyiddeen, L.T. Merce, S. Muttukrishna, L.G. Nardo, S.M. Nelson, H.Y. Ng, B. Popovic-Todorovic, J.M.J. Smeenk, C. Tomás, P.J.Q. Van der Linden, I.A. van Rooij, I.K. Vladimirov, and F.J.M. Broekmans

Analyses: S.L. Broer, M. Dólleman J. van Disseldorp, K.A. Broeze, B.C. Opmeer, P. Bossuyt and M.J.C. Eijkemans

Interpretation of the results and writing of the article: S.L. Broer, M. Dólleman, J. van Disseldorp, K.A. Broeze, B.C. Opmeer, P. Bossuyt, M.J.C. Eijkemans, B.W. Mol and F.J.M. Broekmans

Revision of the article: all authors.

Funding

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

Potential conflict of interests

Prof. F.J.M. Broekmans is a member of the external advisory board for Ferring Pharmaceuticals, Hoofddorp, The Netherlands. He receives no monetary compensation.

All author authors have no potential conflict of interests.

445 **Figure 1. Flowchart of included studies**

446

447

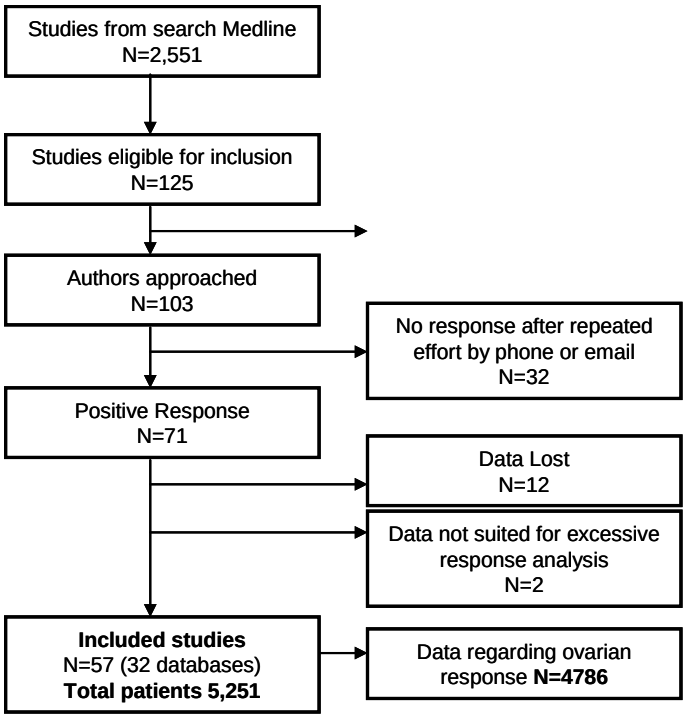
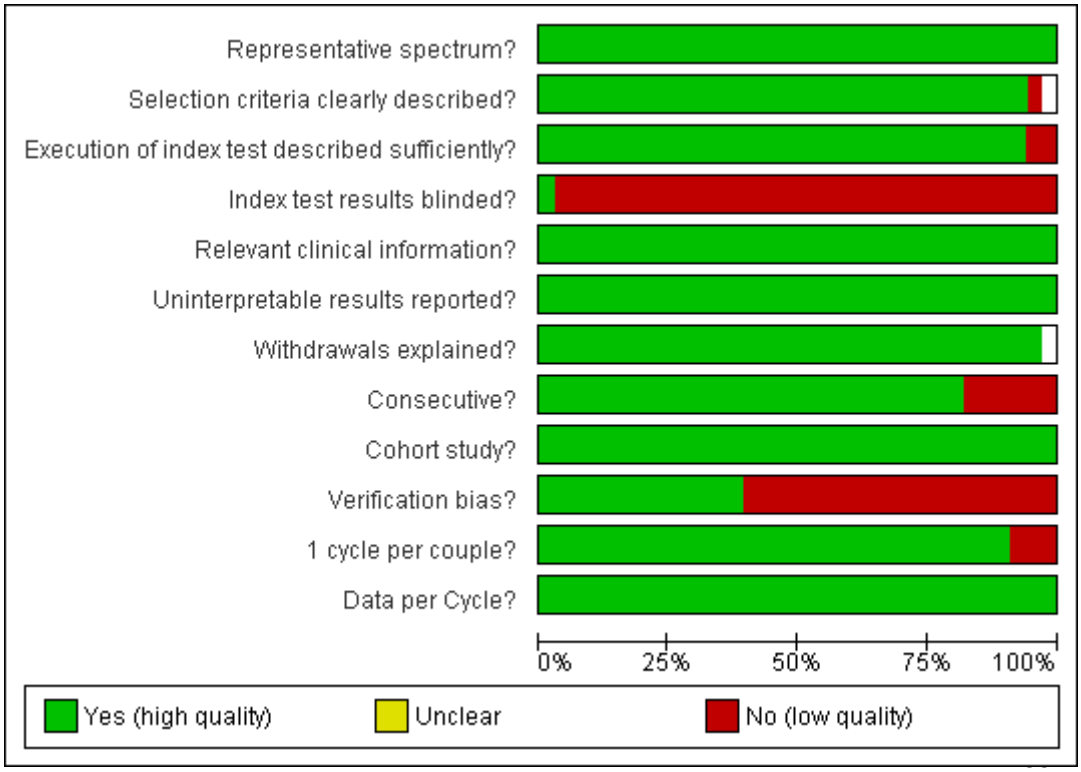


Figure 2: Study characteristics according to QUADAS



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, some studies analyzed more cycles per couple, in which case only the first cycle was analyzed.

467 **Table 1. Baseline characteristics from pooled data.**

	Total population	Excessive Responders	Non-excessive responder	P value
	Mean (5th–95th percentile)	Mean (5th–95th percentile)	Mean (5th–95th percentile)	
Female age (years)	34.4 (26.0-42.0)	32.5 (25.0-39.9)	34.7 (26.0-42.0)	< 0.001
FSH (IU/L)	7.7 (3.8-14.0)	6.4 (3.5-10.1)	8.7 (3.9-16.0)	< 0.001
AFC (number)	12.1 (3.0-25.6)	17.1 (6.0-32.0)	11.0 (3.0-22.0)	< 0.001
AMH (ng/ml)	2.5 (0.1-7.6)	4.8 (1.3-10.2)	2.0 (0.1-5.7)	< 0.001
BMI (kg/m2)	23.6 (18.6-30.1)	23.4 (18.5-29.4)	23.4 (18.6-30.1)	0.943
Duration of subfertility (years)	4.3 (1.3-10.0)	4.3 (1.5-10.0)	4.3 (1.2-10.0)	0.937

468

469 ***Legend.***

470 *Excessive Response definition: > 15 oocytes retrieved. Duration of subfertility: the period from the cessation of*
471 *contraceptive methods or start of unprotected intercourse until the first IVF attempt. Excessive responders N =*
472 *894 (18.7%). Non excessive responders = 3,892.*

473 *AFC, Antral Follicle Count; AMH, Anti-Müllerian Hormone; FSH, Follicle Stimulating Hormone.*

474

Table 2. Univariable and multivariable models of age and ORTs in the prediction of an excessive response

	Three test study group (N= 1,023)				Total study group (N= 4,786)			
	OR	**95% CI**	**P - value**	**Variance-RI**	**OR**	**95% CI**	**P - value**	**Variance-RI**
Univariable models								
Age (per year)	0.89	0.85 - 0.93	<0.001	0.748	0.90	0.88 - 0.91	<0.001	0.543
FSH (per IU/L)	0.76	0.70 - 0.84	<0.001	1.23	0.83	0.80 - 0.86	<0.001	0.551
AFC (per N)	1.18	1.15 - 1.22	<0.001	0.715	1.14	1.12 - 1.16	<0.001	0.605
AMH (per 1.0 ng/ml)	1.61	1.48 - 1.76	<0.001	0.878	1.59	1.49 - 1.70	<0.001	0.680
Multivariable models								
Age and FSH								
Age (per year)	0.91	0.87 - 0.94	<0.001	0.82	0.91	0.89 - 0.93	<0.001	0.497
FSH (per IU/L)	0.79	0.72 - 0.87	<0.001		0.85	0.82 - 0.88	<0.001	
Age and AFC								
Age (per year)	0.93	0.89 - 1.98	0.003	0.769	0.95	0.92 - 0.98	0.001	0.575
AFC (per N)	1.17	1.13 - 1.21	<0.001		1.13	1.11 - 1.15	<0.001	
Age and AMH								
Age (per year)	0.92	0.88 - 0.97	<0.001	0.596	0.92	0.89 - 0.95	<0.001	0.599
AMH (per 1.0 ng/ml)	1.57	1.43 - 1.71	<0.001		1.54	1.44 - 1.64	<0.001	

Legend.

Results of random intercept logistic regression model in the prediction of an excessive response. Multivariable analyses showed that all three ORTs add predictive information to female age alone. P values reflect whether the variable plays a significant role in the model. The column "Variance RI" denotes the estimated variance of the random intercept in the Random intercept logistic model. It's square root is the estimated standard deviation (SD) , and may be interpreted on the logistic scale. A one SD difference between two studies in the population of studies corresponds to an increase in the Odds on the outcome (excessive response) of exp(SD). E.g. the Age and AMH model for excessive response has variance RI = 0.321, so exp(sqrt(0.321))=1.76, is the relative increase in Odds of excessive response corresponding to a difference between two studies in intercept of one SD.
OR (Odds Ratio), 95% CI (95% Confidence Interval).

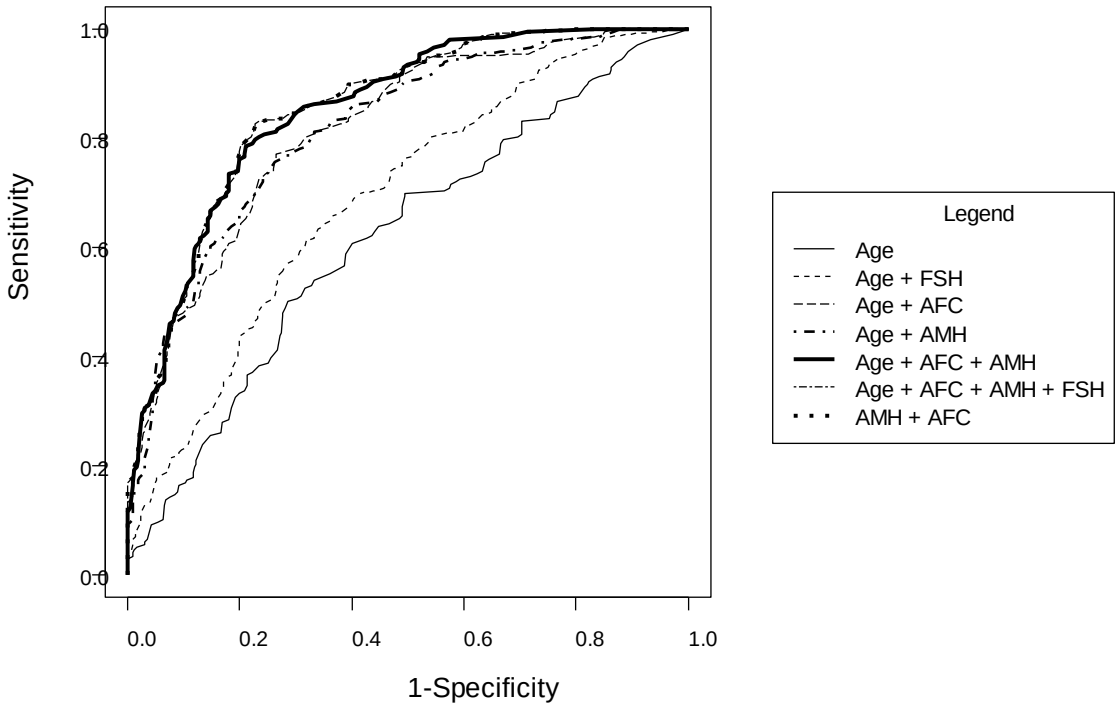
Table 3. AUCs of prediction models of age and ovarian reserve tests for the prediction of an excessive response

	Three test study group				Total study group			
	AUC	**95% CI**	**P value**	**N**	**AUC**	**95% CI**	**P value**	**N**
Univariable analysis								
Age	0.61	0.54 - 0.68	NA	1023	0.61	0.58 - 0.64	NA	4650
FSH	0.66	0.60 - 0.73	0.071	1023	0.64	0.61 - 0.67	0.026	4254
AFC	0.79	0.74 - 0.85	< 0.001	1023	0.73	0.69 - 0.77	< 0.001	2524
AMH	0.81	0.76 - 0.87	< 0.001	1023	0.82	0.77 - 0.86	< 0.001	1890
Multivariable analysis								
Age & FSH	0.68	0.62 - 0.75	< 0.001	1023	0.67	0.64 - 0.71	< 0.001	4254
Age & AFC	0.81	0.76 - 0.87	< 0.001	1023	0.75	0.71 - 0.79	< 0.001	2524
Age & AMH	0.81	0.76 - 0.87	< 0.001	1023	0.81	0.77 - 0.85	< 0.001	1890
Age & AMH & AFC	0.85	0.80 - 0.90	< 0.001	1023	0.85	0.80 - 0.90	< 0.001	1024
Age & AMH & AFC & FSH	0.85	0.80 - 0.90	< 0.001	1023	0.85	0.80 - 0.90	< 0.001	1023
AMH & AFC	0.85	0.80 - 0.90	< 0.001	1023	0.85	0.80 - 0.90	< 0.001	1024

Legend.

The Area Under the Curve (AUC) of the univariable and multivariable models of age or ORTs in the prediction of an excessive response are shown. In the univariable analysis it is shown that both AMH and AFC have a high accuracy, while FSH only has a moderate accuracy. In the multivariable models the added value to the AUC of an ORT on female age is shown, the p value indicates whether this added value is significant in comparison to the model based on age alone. Adding any of the ORTs shows a significant rise in the AUC. Moreover, the added value of adding several ORTs to female age is shown. The model including age, AFC and AMH reached the maximum predictive power. Addition of FSH to this model did not improve the predictive accuracy (P = 0.725). However, a model with AMH and AFC alone has a comparable AUC.

Figure 2 ROC curves of age and ORTs in the prediction of an excessive response



Legend.

The ROC curves of age and age combined with a single or more ORTs are depicted. The ROC curves for 'Age + AMH', 'Age + AFC', 'Age + AMH + AFC' and 'Age + AMH + AFC + FSH' run toward the upper left corner of the ROC space, indicating a good capacity to discriminate between normal and excessive responders at certain cut-off levels.NB ROC curves in the three-test study group (N = 1023). AFC, Antral Follicle Count; AMH, Anti-Müllerian Hormone; FSH, Follicle Stimulating Hormone; ORT, Ovarian Reserve Test; ROC, receiver-operating characteristic.

519 **Table 4. Results of the ROC regression analysis.**

520

	Coefficient	*95% CI*	*P-value*
		Age	
FSH	-0.029	-0.051 - -0.006	**0.010**
AFC	0.032	0.006 - 0.056	**0.010**
AMH	-0.021	-0.049 - 0.005	0.139
		BMI	
FSH	0.026	-0.024 - 0.070	0.267
AFC	-0.009	-0.048 - 0.033	0.674
AMH	0.019	-0.024 - 0.056	0.363
		Duration	
FSH	0.018	-0.044 - 0.078	0.569
AFC	0.047	-0.022 - 0.112	0.177
AMH	-0.041	-0.113 - 0.026	0.246

521

522 ***Legend.***

523 *ROC regression analysis showing the effect of the patient characteristics on the ROC curve of the*
524 *ovarian reserve tests in the prediction of an excessive ovarian response.*

525 *Bold = significant influence of the patient characteristics on the discriminatory capacity of the ovarian reserve*
526 *test in the prediction of an excessive response. AFC = Antral Follicle Count; AMH = Anti-Müllerian Hormone;*
527 *FSH = Follicle Stimulating Hormone; Duration= Duration of subfertility.*

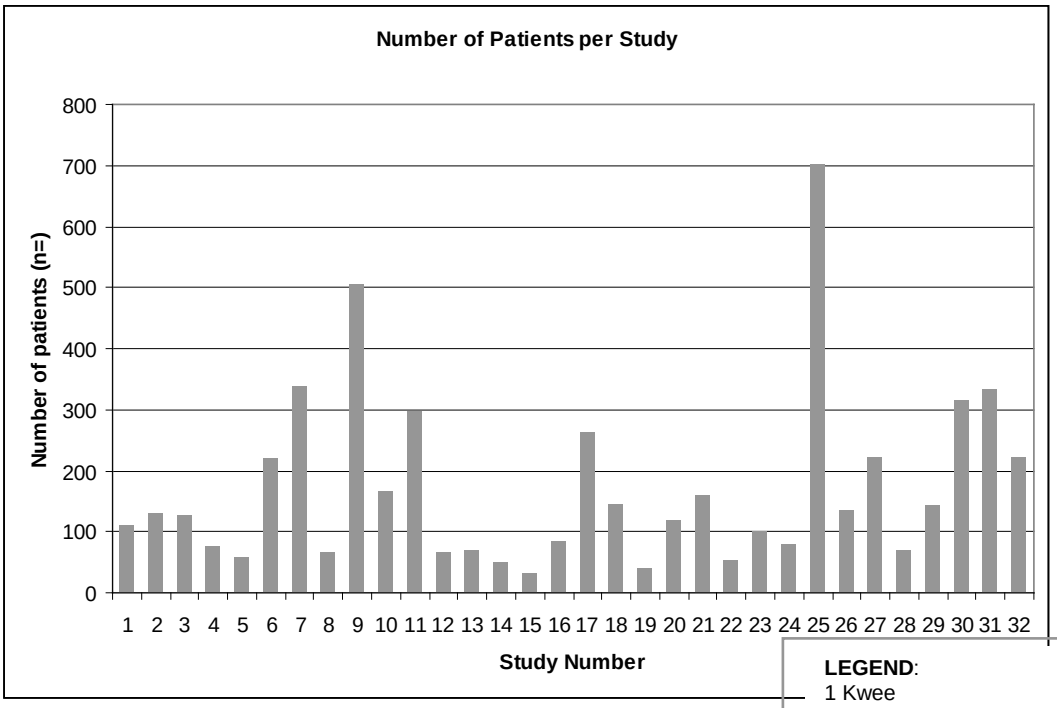
528

529

ADDENDUM

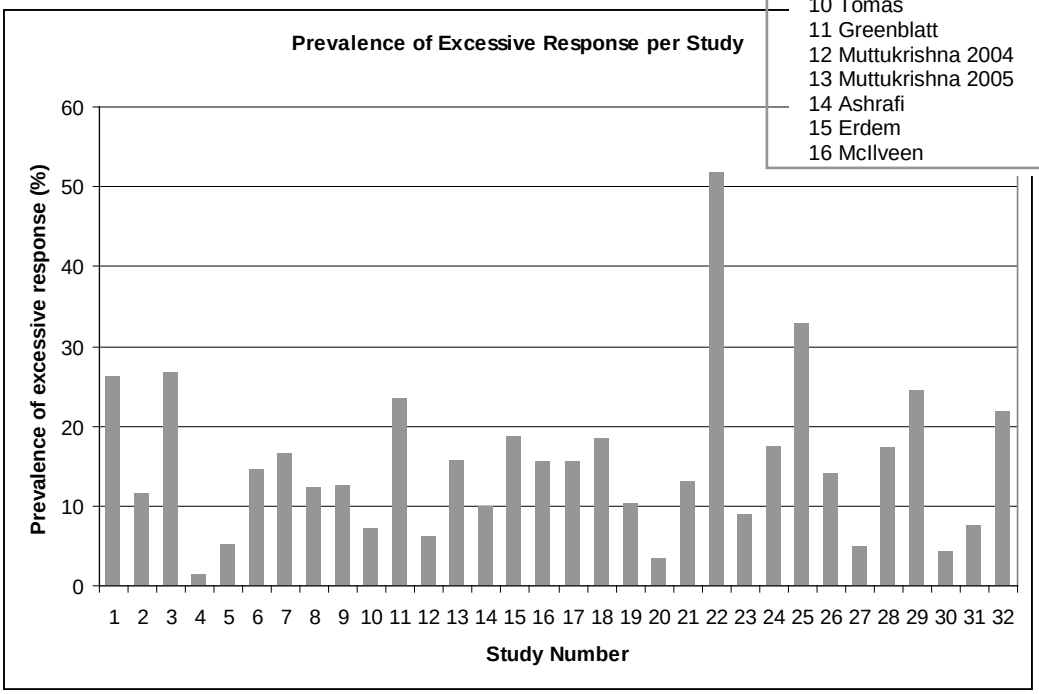
Figure A-1. Baseline characteristics of the included studies

A. Number of patients per study

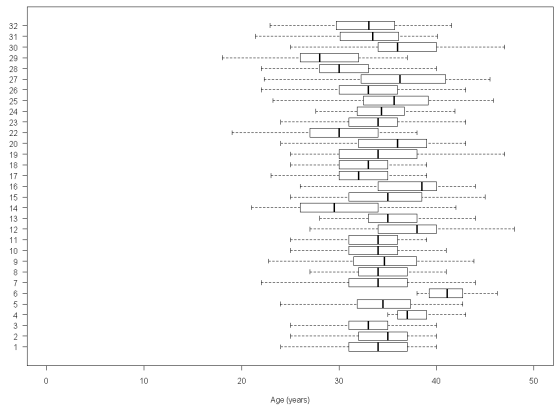


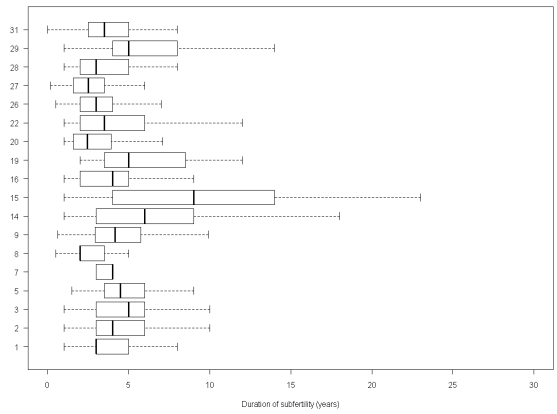
LEGEND:
1 Kwee
2 Ng 2000
3 Ng 2005
4 Caroppo
5 Anderson
6 Klinkert
7 Nelson
8 Merce
9 Bancsi
10 Tomàs
11 Greenblatt
12 Muttukrishna 2004
13 Muttukrishna 2005
14 Ashrafi
15 Erdem
16 McIlveen
17 Popovic 2003a
18 Popovic 2003b
19 Vladimirov
20 La Marca
21 van der Linden
22 Eldar-Geva
23 Jayaprakasan
24 Smeenk 2007
25 Copperman
26 Ebner
27 van Rooij
28 Freour
29 Aflatoonian
30 Gnoth
31 Nardo **unpublished*
32 Nardo 2008

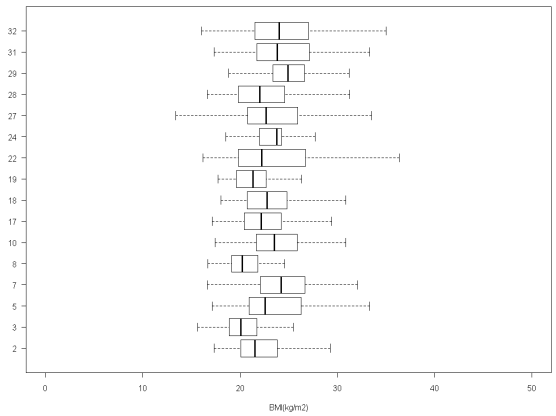
B. Incidence of an excessive response per study



C. Patient Characteristics



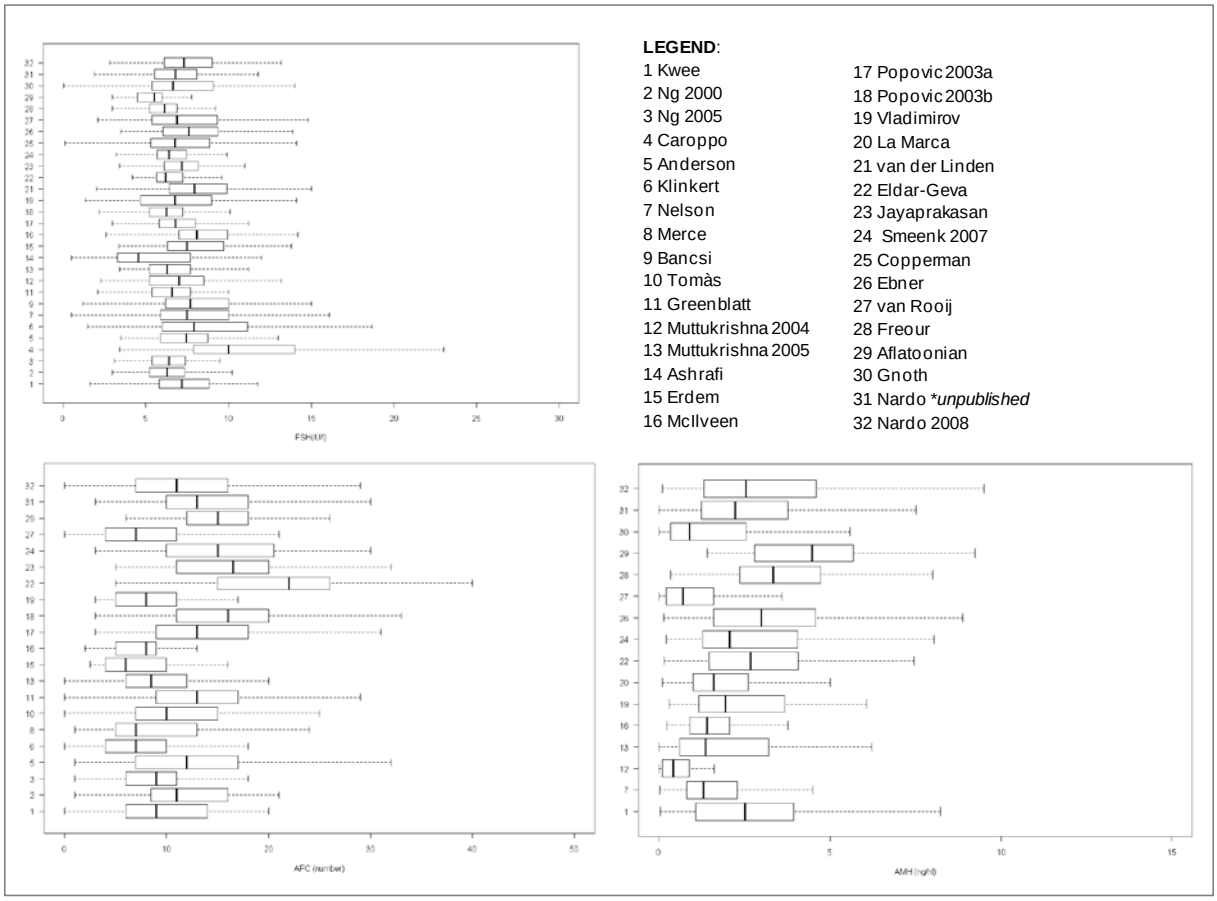




LEGEND:

1 Kwee
2 Ng 2000
3 Ng 2005
4 Caroppo
5 Anderson
6 Klinkert
7 Nelson
8 Merce
9 Bancsi
10 Tomàs
11 Greenblatt
12 Muttukrishna 2004
13 Muttukrishna 2005
14 Ashrafi
15 Erdem
16 McIlveen
17 Popovic 2003a
18 Popovic 2003b
19 Vladimirov
20 La Marca
21 van der Linden
22 Eldar-Geva
23 Jayaprakasan
24 Smeenk 2007
25 Copperman
26 Ebner
27 van Rooij
28 Freour
29 Aflatoonian
30 Gnoth
31 Nardo **unpublished*
32 Nardo 2008

D. Ovarian Reserve Tests



Legend:

A. The number of patients per study are demonstrated
B. The prevalence of an excessive response per study is demonstrated
C. For each individual study the mean, 5th and 95th percentile of the patient characteristics female age, BMI and duration of subfertility are shown.
D. For each individual study the mean, 5th and 95th percentile of ovarian reserve tests FSH, AFC and AMH are shown.

Table A-1. AUCs of the included studies in the prediction of an excessive response

Study	FSH		AFC		AMH	
	AUC	*N*	*AUC*	*N*	*AUC*	*N*
Aflatoonian	0.60 (0.50-0.69)	143	0.96 (0.93-0.99)	143	0.94 (0.90-0.98)	143
Anderson	0.92 (0.99-1.00)	46	0.61(0.67-0.85)	46	NA	
Ashrafi	0.59 (0.31-0.87)	50	NA		NA	
Bancsi	0.61(0.54-0.68)	505	NA		NA	
Caroppo	0.81(0.72-0.90)	76	NA		NA	
Copperman	0.65 (0.60-0.69)	570	NA		NA	
Ebner	0.61 (0.46-0.75)	127	NA		0.82 (0.74-0.90)	135
Eldar-Geva	0.71(0.57-0.85)	52	0.88 (0.75-1.00)	36	0.75 (0.62-0.88)	54
Erdem	0.77 (0.57-0.97)	24	0.85 (0.70-1.00)	24	NA	
Freour	0.58 (0.41-0.73)	62	NA		0.70 (0.55-0.86)	64
Gnoth	0.64 (0.51-0.78)	122	NA		0.87 (0.79-0.95)	134
Greenblatt	0.67(0.59-0.74)	261	0.69 (0.61-0.77)	223	NA	
Jayaprakasan	0.74(0.57-0.91)	100	0.82 (0.70-0.95)	100	NA	
Klinkert	0.42 (0.30-0.55)	212	0.45 (0.33-0.57)	221	NA	
Kwee	0.79 (0.70-0.88)	109	0.87 (0.82-0.96)	109	0.84 (0.76-0.92)	105
La Marca	NA		NA		0.90 (0.76-1.00)	118
McIlveen	No >15	71	No >15	71	No >15	
Merce	NA		0.62 (0.42-0.83)	65	NA	
Muttukrishna 1	0.81 (0.59-1.00)	66	NA		0.92 (0.83-1.00)	66
Muttukrishna 2	0.67 (0.52-0.82)	68	0.84 (0.73-0.94)	68	0.73 (0.56-0.91)	68
Nardo 1	0.65 (0.53-0.77)	135	0.71(0.59-0.83)	123	0.74 (0.64-0.83)	135
Nardo 2	0.68 (0.59-0.77)	145	0.71(0.63-0.80)	145	0.79 (0.72-0.87)	145
Nelson	0.64 (0.58-0.71)	338	NA		0.88 (0.82-0.91)	319
Ng 1	0.70 (0.56-0.83)	131	0.80 (0.70-0.90)	131	NA	
Ng 2	0.72 (0.56-0.83)	109	0.77 (0.68-0.85)	127	NA	
Popovic 1	0.62 (0.54-0.71)	256	0.71(0.63-0.80)	256	NA	
Popovic 2	0.62 (0.50-0.73)	143	0.76 (0.67-0.86)	143	NA	
Smeenk 1	0.54 (0.40-0.68)	80	0.66 (0.5300.79)	80	0.71 (0.57-0.84)	80
Smeenk 2	NA		NA		NA	
Tomas	NA		0.82 (0.72-0.91)	160	NA	
Van Rooij	0.68 (0.58-0.79)	215	0.86 (0.79-0.93)	215	0.87 (0.77-0.97)	215
Van der Linden	0.82 (0.72-0.92)	124	NA		NA	
Vladimirov 2	0.67 (0.48-0.87)	39	0.74 (0.52-0.97)	39	0.80 (0.67-0.93)	39

Table A-2. Univariable and multivariable models of patient characteristics in the prediction of an excessive response

Excessive Response Prediction						
	Three test study group			Total study group		
	OR	95% CI	P - value	OR	95% CI	P - value
Univariable models						
Age	0.89	0.85 - 0.93	< 0.001	0.90	0.88 - 0.91	< 0.001
BMI	0.98	0.93 - 1.03	0.405	1.00	0.97 - 1.03	0.954
Duration	0.98	0.90 - 1.06	0.555	0.97	0.92 - 1.01	0.156
Multivariable models						
Age and BMI						
Age	0.91	0.87 - 0.95	< 0.001	0.9	0.87 - 0.93	< 0.001
BMI	0.99	0.93 - 1.04	0.616	1.00	0.97 - 1.04	0.976
Age and duration						
Age	0.90	0.85 - 0.94	< 0.001	0.89	0.86 - 0.91	< 0.001
Duration	1.01	0.93 - 1.10	0.750	1.00	0.95 - 1.05	0.956

Legend.

OR = Odds Ratio, 95%CI = 95% Confidence Interval. Duration = duration of subfertility.

Reference List

1. Aflatoonian A, Oskouian H, Ahmadi S, and Oskouian L (2009) Prediction of high ovarian response to controlled ovarian hyperstimulation: anti-Mullerian hormone versus small antral follicle count (2-6 mm). J Assist Reprod Genet, 26, 319-325.

2. Anckaert E, Smitz J, Schiettecatte J, Klein BM, and Arce JC (2012) The value of anti-Mullerian hormone measurement in the long GnRH agonist protocol: association with ovarian response and gonadotrophin-dose adjustments. Hum Reprod, 27, 1829-1839.

3. Andersen AN, Witjes H, Gordon K, and Mannaerts B (2011) Predictive factors of ovarian response and clinical outcome after IVF/ICSI following a rFSH/GnRH antagonist protocol with or without oral contraceptive pre-treatment. Hum Reprod, 26, 3413-3423.

4. Ashrafi M, Madani T, Tehranian AS, and Malekzadeh F (2005) Follicle stimulating hormone as a predictor of ovarian response in women undergoing controlled ovarian hyperstimulation for IVF. Int J Gynaecol Obstet, 91, 53-57.

5. Baart EB, Martini E, van dB, I, Macklon NS, Galjaard RJ, Fauser BC, and van Opstal D (2006) Preimplantation genetic screening reveals a high incidence of aneuploidy and mosaicism in embryos from young women undergoing IVF. Hum Reprod, 21, 223-233.

6. Bancsi LF, Huijs AM, Den Ouden CT, Broekmans FJ, Looman CW, Blankenstein MA, and te Velde ER (2000) Basal follicle-stimulating hormone levels are of limited value in predicting ongoing pregnancy rates after in vitro fertilization. Fertil Steril, 73, 552-557.

7. Broekmans FJ, Kwee J, Hendriks DJ, Mol BW, and Lambalk CB (2006) A systematic review of tests predicting ovarian reserve and IVF outcome. Hum Reprod Update, 12, 685-718.

8. Broer SL (2011) Assessment of Current and Future Ovarian Reserve Status. In Gildeprint Drukkerijen, Enschede, The Netherlands.

9. Broer SL, Dolleman M, Opmeer BC, Fauser BC, Mol BW, and Broekmans FJ (2011) AMH and AFC as predictors of excessive response in controlled ovarian hyperstimulation: a meta-analysis. Hum Reprod Update, 17, 46-54.

10. Broer SL, Eijkemans MJ, Scheffer GJ, van R, I, de VA, Themmen AP, Laven JS, de Jong FH, te Velde ER, Fauser BC et al (2011) Anti-mullerian hormone predicts menopause: a long-term follow-up study in normoovulatory women. J Clin Endocrinol Metab, 96, 2532-2539.

11. Broer SL, Mol B, Dolleman M, Fauser BC, and Broekmans FJ (2010) The role of anti-Mullerian hormone assessment in assisted reproductive technology outcome. Curr Opin Obstet Gynecol, 22, 193-201.

12. Broer SL, Mol BW, Hendriks D, and Broekmans FJ (2009) The role of antimullerian hormone in prediction of outcome after IVF: comparison with the antral follicle count. Fertil Steril, 91, 705-714.

13. Broeze KA, Opmeer BC, Bachmann LM, Broekmans FJ, Bossuyt PM, Coppus SF, Johnson NP, Khan KS, ter RG, van d, V et al (2009) Individual patient data meta-analysis of diagnostic and prognostic studies in obstetrics, gynaecology and reproductive medicine. BMC Med Res Methodol, 9, 22.

14. Broeze KA, Opmeer BC, Van GN, Coppus SF, Collins JA, Den Hartog JE, van der Linden PJ, Marianowski P, Ng EH, Van der Steeg JW et al (2011) Are patient characteristics associated with the accuracy of hysterosalpingography in diagnosing tubal pathology? An individual patient data meta-analysis. Hum Reprod Update, 17, 293-300.

15. Caroppo E, Matteo M, Schonauer LM, Vizziello G, Pasquadibisceglie A, Vitti A, and D'Amato G (2006) Basal FSH concentration as a predictor of IVF outcome in older women undergoing stimulation with GnRH antagonist. Reprod Biomed Online, 13, 815-820.

16. Delvigne A and Rozenberg S (2002) Epidemiology and prevention of ovarian hyperstimulation syndrome (OHSS): a review. Hum Reprod Update, 8, 559-577.

17. Eldar-Geva T, Ben Chetrit A, Spitz IM, Rabinowitz R, Markowitz E, Mimoni T, Gal M, Zylber-Haran E, and Margalioth EJ (2005a) Dynamic assays of inhibin B, anti-Mullerian hormone and estradiol following FSH stimulation and ovarian ultrasonography as predictors of IVF outcome. Hum Reprod, 20, 3178-3183.

18. Eldar-Geva T, Ben Chetrit A, Spitz IM, Rabinowitz R, Markowitz E, Mimoni T, Gal M, Zylber-Haran E, and Margalioth EJ (2005b) Dynamic assays of inhibin B, anti-Mullerian hormone and estradiol following FSH stimulation and ovarian ultrasonography as predictors of IVF outcome. Hum Reprod, 20, 3178-3183.

19. Erdem M, Erdem A, Gursoy R, and Biberoglu K (2004) Comparison of basal and clomiphene citrate induced FSH and inhibin B, ovarian volume and antral follicle counts as ovarian reserve tests and predictors of poor ovarian response in IVF. J Assist Reprod Genet, 21, 37-45.

20. Fauser BC, Diedrich K, and Devroey P (2008) Predictors of ovarian response: progress towards individualized treatment in ovulation induction and ovarian stimulation. Hum Reprod Update, 14, 1-14.

21. Freour T, Mirallie S, Bach-Ngohou K, Denis M, Barriere P, and Masson D (2007) Measurement of serum anti-Mullerian hormone by Beckman Coulter ELISA and DSL ELISA: comparison and relevance in assisted reproduction technology (ART). Clin Chim Acta, 375, 162-164.

22. Gnoth C, Schuring AN, Friol K, Tigges J, Mallmann P, and Godehardt E (2008) Relevance of anti-Mullerian hormone measurement in a routine IVF program. Hum Reprod, 23, 1359-1365.

23. Heijnen EM, Eijkemans MJ, De KC, Polinder S, Beckers NG, Klinkert ER, Broekmans FJ, Passchier J, te Velde ER, Macklon NS et al (2007) A mild treatment strategy for in-vitro fertilisation: a randomised non-inferiority trial. Lancet, 369, 743-749.

24. Ho HY, Lee RK, Lin MH, and Hwu YM (2003) Estradiol level on day 9 as a predictor of risk for ovarian hyperresponse during controlled ovarian hyperstimulation. J Assist Reprod Genet, 20, 222-226.

25. Janes H, Longton G, and Pepe M (2009) Accommodating Covariates in ROC Analysis. Stata J, 9, 17-39.

26. Jayaprakasan K, Hilwah N, Kendall NR, Hopkisson JF, Campbell BK, Johnson IR, and Raine-Fenning NJ (2007) Does 3D ultrasound offer any advantage in the pretreatment assessment of ovarian reserve and prediction of outcome after assisted reproduction treatment? Hum Reprod, 22, 1932-1941.

27. Jayaprakasan K, Hopkisson J, Campbell B, Johnson I, Thornton J, and Raine-Fenning N (2010) A randomised controlled trial of 300 versus 225 IU recombinant FSH for ovarian stimulation in predicted normal responders by antral follicle count. BJOG, 117, 853-862.

28. Klinkert ER, Broekmans FJ, Looman CW, Habbema JD, and te Velde ER (2005) The antral follicle count is a better marker than basal follicle-stimulating hormone for the selection of older patients with acceptable pregnancy prospects after in vitro fertilization. Fertil Steril, 83, 811-814.

29. Kwee J, Elting MW, Schats R, Bezemer PD, Lambalk CB, and Schoemaker J (2003) Comparison of endocrine tests with respect to their predictive value on the outcome of ovarian hyperstimulation in IVF treatment: results of a prospective randomized study. Hum Reprod, 18, 1422-1427.

30. La Marca A, Giulini S, Tirelli A, Bertucci E, Marsella T, Xella S, and Volpe A (2007) Anti-Mullerian hormone measurement on any day of the menstrual cycle strongly predicts ovarian response in assisted reproductive technology. Hum Reprod, 22, 766-771.

31. Lee S, Ozkavukcu S, Heytens E, Moy F, Alappat RM, and Oktay K (2011) Anti-Mullerian hormone and antral follicle count as predictors for embryo/oocyte cryopreservation cycle outcomes in breast cancer patients stimulated with letrozole and follicle stimulating hormone. J Assist Reprod Genet, 28, 651-656.

32. Liu KE and Greenblatt EM (2008) Elevated day 3 follicle-stimulating hormone/luteinizing hormone ratio >or= 2 is associated with higher rates of cancellation in in vitro fertilization-embryo transfer cycles. Fertil Steril, 90, 297-301.

33. Luna M, Grunfeld L, Mukherjee T, Sandler B, and Copperman AB (2007) 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, 87, 782-787.

34. McIlveen M, Skull JD, and Ledger WL (2007) Evaluation of the utility of multiple endocrine and ultrasound measures of ovarian reserve in the prediction of cycle cancellation in a high-risk IVF population. Hum Reprod, 22, 778-785.

35. Merce LT, Barco MJ, Bau S, and Troyano JM (2007) Prediction of ovarian response and IVF/ICSI outcome by three-dimensional ultrasonography and power Doppler angiography. Eur J Obstet Gynecol Reprod Biol, 132, 93-100.

36. Muttukrishna S, McGarrigle H, Wakim R, Khadum I, Ranieri DM, and Serhal P (2005) Antral follicle count, anti-mullerian hormone and inhibin B: predictors of ovarian response in assisted reproductive technology? BJOG, 112, 1384-1390.

37. Muttukrishna S, Suharjono H, McGarrigle H, and Sathanandan M (2004) Inhibin B and anti-Mullerian hormone: markers of ovarian response in IVF/ICSI patients? BJOG, 111, 1248-1253.

38. Nakhuda GS, Chu MC, Wang JG, Sauer MV, and Lobo RA (2006) Elevated serum mullerian-inhibiting substance may be a marker for ovarian hyperstimulation syndrome in normal women undergoing in vitro fertilization. Fertil Steril, 85, 1541-1543.

39. Nakhuda GS, Douglas NC, Thornton MH, Guarnaccia MM, Lobo R, and Sauer MV (2011) Anti-Mullerian hormone testing is useful for individualization of stimulation protocols in oocyte donors. Reprod Biomed Online, 22 Suppl 1, S88-S93.

40. Nardo LG, Gelbaya TA, Wilkinson H, Roberts SA, Yates A, Pemberton P, and Laing I (2009) Circulating basal anti-Mullerian hormone levels as predictor of ovarian response in women undergoing ovarian stimulation for in vitro fertilization. Fertil Steril, 92, 1586-1593.

41. Nelson SM, Yates RW, and Fleming R (2007) Serum anti-Mullerian hormone and FSH: prediction of live birth and extremes of response in stimulated cycles--implications for individualization of therapy. Hum Reprod, 22, 2414-2421.

42. Ng EH, Chan CC, Tang OS, and Ho PC (2005) Antral follicle count and FSH concentration after clomiphene citrate challenge test in the prediction of ovarian response during IVF treatment. Hum Reprod, 20, 1647-1654.

43. Ng EH, Tang OS, and Ho PC (2000) The significance of the number of antral follicles prior to stimulation in predicting ovarian responses in an IVF programme. Hum Reprod, 15, 1937-1942.

44. Olivennes F, Howies CM, Borini A, Germond M, Trew G, Wikland M, Zegers-Hochschild F, Saunders H, and Alam V (2011) Individualizing FSH dose for assisted reproduction using a novel algorithm: the CONSORT study. Reprod Biomed Online, 22 Suppl 1, S73-S82.

45. Pepe M, Longton G, and Janes H (2009) Estimation and Comparison of Receiver Operating Characteristic Curves. Stata J, 9, 1.

46. Popovic-Todorovic B, Loft A, Bredkjaeer HE, Bangsboll S, Nielsen IK, and Andersen AN (2003b) A prospective randomized clinical trial comparing an individual dose of recombinant FSH based on predictive factors versus a 'standard' dose of 150 IU/day in 'standard' patients undergoing IVF/ICSI treatment. Hum Reprod, 18, 2275-2282.

47. Popovic-Todorovic B, Loft A, Bredkjaeer HE, Bangsboll S, Nielsen IK, and Andersen AN (2003a) A prospective randomized clinical trial comparing an individual dose of recombinant FSH based on predictive factors versus a 'standard' dose of 150 IU/day in 'standard' patients undergoing IVF/ICSI treatment. Hum Reprod, 18, 2275-2282.

48. Popovic-Todorovic B, Loft A, Lindhard A, Bangsboll S, Andersson AM, and Andersen AN (2003d) A prospective study of predictive factors of ovarian response in 'standard' IVF/ICSI patients treated with recombinant FSH. A suggestion for a recombinant FSH dosage normogram. Hum Reprod, 18, 781-787.

49. Popovic-Todorovic B, Loft A, Lindhard A, Bangsboll S, Andersson AM, and Andersen AN (2003c) A prospective study of predictive factors of ovarian response in 'standard' IVF/ICSI patients treated with recombinant FSH. A suggestion for a recombinant FSH dosage normogram. Hum Reprod, 18, 781-787.

50. Riggs R, Kimble T, Oehninger S, Bocca S, Zhao Y, Leader B, and Stadtmauer L (2011) Anti-Mullerian hormone serum levels predict response to controlled ovarian hyperstimulation but not embryo quality or pregnancy outcome in oocyte donation. Fertil Steril, 95, 410-412.

51. Riggs RM, Duran EH, Baker MW, Kimble TD, Hobeika E, Yin L, Matos-Bodden L, Leader B, and Stadtmauer L (2008) Assessment of ovarian reserve with anti-Mullerian hormone: a comparison of the predictive value of anti-Mullerian hormone, follicle-stimulating hormone, inhibin B, and age. Am J Obstet Gynecol, 199, 202-208.

52. Riley RD, Dodd SR, Craig JV, Thompson JR, and Williamson PR (2008) Meta-analysis of diagnostic test studies using individual patient data and aggregate data. Stat Med, 27, 6111-6136.

53. Smeenk JM, Stolwijk AM, Kremer JA, and Braat DD (2000) External validation of the templeton model for predicting success after IVF. Hum Reprod, 15, 1065-1068.

54. Smeenk JM, Sweep FC, Zielhuis GA, Kremer JA, Thomas CM, and Braat DD (2007) Antimullerian hormone predicts ovarian responsiveness, but not embryo quality or pregnancy, after in vitro fertilization or intracyoplasmic sperm injection. Fertil Steril, 87, 223-226.

55. Steinkampf MP, Hammond KR, Nichols JE, and Slayden SH (2003) Effect of obesity on recombinant follicle-stimulating hormone absorption: subcutaneous versus intramuscular administration. Fertil Steril, 80, 99-102.

56. Sterrenburg MD, Veltman-Verhulst SM, Eijkemans MJ, Hughes EG, Macklon NS, Broekmans FJ, and Fauser BC (2011) Clinical outcomes in relation to the daily dose of recombinant follicle-stimulating hormone for ovarian stimulation in in vitro fertilization in presumed normal responders younger than 39 years: a meta-analysis. Hum Reprod Update, 17, 184-196.

57. Tomas-C, Nuojua-Huttunen-, and Martikainen-H (1997) Pretreatment transvaginal ultrasound examination predicts ovarian responsiveness to gonadotrophins in in-vitro fertilization. Hum-Reprod, 12, 220-223.

58. van der Gaast MH, Eijkemans MJ, van der Net JB, de Boer EJ, Burger CW, van Leeuwen FE, Fauser BC, and Macklon NS (2006) Optimum number of oocytes for a successful first IVF treatment cycle. Reprod Biomed Online, 13, 476-480.

59. Van der Meer M, Hompes PG, De Boer JA, Schats R, and Schoemaker J (1998) Cohort size rather than follicle-stimulating hormone threshold level determines ovarian sensitivity in polycystic ovary syndrome. J Clin Endocrinol Metab, 83, 423-426.

60. van Rooij I, Broekmans FJ, te Velde ER, Fauser BC, Bancsi LF, de Jong FH, and Themmen AP (2002a) Serum anti-Mullerian hormone levels: a novel measure of ovarian reserve. Hum Reprod, 17, 3065-3071.

61. van Rooij IA, Broekmans FJ, te Velde ER, Fauser BC, Bancsi LF, Jong FH, and Themmen AP (2002b) Serum anti-Mullerian hormone levels: a novel measure of ovarian reserve. Hum Reprod, 17, 3065-3071.

62. van Swieten EC, Leeuw-Harmsen L, Badings EA, and van der Linden PJ (2005) Obesity and Clomiphene Challenge Test as predictors of outcome of in vitro fertilization and intracytoplasmic sperm injection. Gynecol Obstet Invest, 59, 220-224.

63. van Tilborg TC, Eijkemans MJ, Laven JS, Koks CA, de Bruin JP, Scheffer GJ, van Golde RJ, Fleischer K, Hoek A, Nap AW et al (2012) The OPTIMIST study: optimisation of cost effectiveness through individualised FSH stimulation dosages for IVF treatment. A randomised controlled trial. BMC Womens Health, 12, 29.

64. Verberg MF, Eijkemans MJ, Macklon NS, Heijnen EM, Baart EB, Hohmann FP, Fauser BC, and Broekmans FJ (2009) The clinical significance of the retrieval of a low number of oocytes following mild ovarian stimulation for IVF: a meta-analysis. Hum Reprod Update, 15, 5-12.

65. Vladimirov IK, Tacheva DM, and Kalinov KB (2004) Mean ovarian diameter (MOD) as a predictor of poor ovarian response. J Assist Reprod Genet, 21, 73-77.

66. Vladimirov IK, Tacheva DM, Kalinov KB, Ivanova AV, and Blagoeva VD (2005) Prognostic value of some ovarian reserve tests in poor responders. Arch Gynecol Obstet, 272, 74-79.

67. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, and Bossuyt PM (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med, 155, 529-536.

68. Yong PY, Baird DT, Thong KJ, McNeilly AS, and Anderson RA (2003) Prospective analysis of the relationships between the ovarian follicle cohort and basal FSH concentration, the inhibin response to exogenous FSH and ovarian follicle number at different stages of the normal menstrual cycle and after pituitary down-regulation. Hum Reprod, 18, 35-44.