Human Reproduction Update, Vol.12, No.6 pp. 685–718, 2006 doi:10.1093/humupd/dml034
Advance Access publication August 4, 2006

A systematic review of tests predicting ovarian reserve and IVF outcome

F.J.Broekmans[1], J.Kwee[2], D.J.Hendriks[1], B.W.Mol[3] and C.B.Lambalk[2,4]

[1]Department of Reproductive Medicine, University Medical Centre Utrecht, Utrecht, [2]Division of Reproductive Endocrinology and Fertility and the IVF Centre, Department of Obstetrics and Gynaecology, Vrije Universiteit Medical Centre and [3]Centre for Reproductive Medicine, Department of Obstetrics and Gynecology, Academic Medical Centre, Amsterdam, The Netherlands

[4]To whom correspondence should be addressed at: Department of Reproductive Medicine, Vrije Universiteit Medical Center (VUmc), PO Box 70057, 1007 MB, Amsterdam, The Netherlands. E-mail: cb.lambalk@vumc.nl

The age-related decline of the success in IVF is largely attributable to a progressive decline of ovarian oocyte quality and quantity. Over the past two decades, a number of so-called ovarian reserve tests (ORTs) have been designed to determine oocyte reserve and quality and have been evaluated for their ability to predict the outcome of IVF in terms of oocyte yield and occurrence of pregnancy. Many of these tests have become part of the routine diagnostic procedure for infertility patients who undergo assisted reproductive techniques. The unifying goals are traditionally to find out how a patient will respond to stimulation and what are their chances of pregnancy. Evidence-based medicine has progressively developed as the standard approach for many diagnostic procedures and treatment options in the field of reproductive medicine. We here provide the first comprehensive systematic literature review, including an *a priori* protocolized information retrieval on all currently available and applied tests, namely early-follicular-phase blood values of FSH, estradiol, inhibin B and anti-Müllerian hormone (AMH), the antral follicle count (AFC), the ovarian volume (OVVOL) and the ovarian blood flow, and furthermore the Clomiphene Citrate Challenge Test (CCCT), the exogenous FSH ORT (EFORT) and the gonadotrophin agonist stimulation test (GAST), all as measures to predict ovarian response and chance of pregnancy. We provide, where possible, an integrated receiver operating characteristic (ROC) analysis and curve of all individual evaluated published papers of each test, as well as a formal judgement upon the clinical value. Our analysis shows that the ORTs known to date have only modest-to-poor predictive properties and are therefore far from suitable for relevant clinical use. Accuracy of testing for the occurrence of poor ovarian response to hyperstimulation appears to be modest. Whether the a *priori* identification of actual poor responders in the first IVF cycle has any prognostic value for their chances of conception in the course of a series of IVF cycles remains to be established. The accuracy of predicting the occurrence of pregnancy is very limited. If a high threshold is used, to prevent couples from wrongly being refused IVF, a very small minority of IVF-indicated cases (~3%) are identified as having unfavourable prospects in an IVF treatment cycle. Although mostly inexpensive and not very demanding, the use of any ORT for outcome prediction cannot be supported. As poor ovarian response will provide some information on OR status, especially if the stimulation is maximal, entering the first cycle of IVF without any prior testing seems to be the preferable strategy.

Key words: IVF/ICSI outcome/ovarian reserve/ovarian stimulation

Introduction

In Western societies the introduction in the 1960s of reliable methods of contraception has led to the birth of fewer children per family. Driven by increasing levels of female education, a growing participation in labour force and career demands, postponement of childbearing has been a secondary consequence of the so-called sexual revolution (Leridon, 1998). These societal changes in family planning have caused a significant increase in the incidence of unwanted infertility due to female reproductive ageing (Weinstein *et al*., 1993; Abma *et al*., 1997; Ventura *et al*., 2001).

From studies on natural populations in which no consistent methods of birth control are applied, it has been shown that natural fertility starts to decline after the age of 30, accelerates in the mid-30s and will lead to sterility at a mean age of 41 (Spira, 1988; Wood, 1989; te Velde and Pearson, 2002) (Figure 1). The reduction in female fertility can also be shown from contemporary population studies. The chance of not conceiving a first child within one year

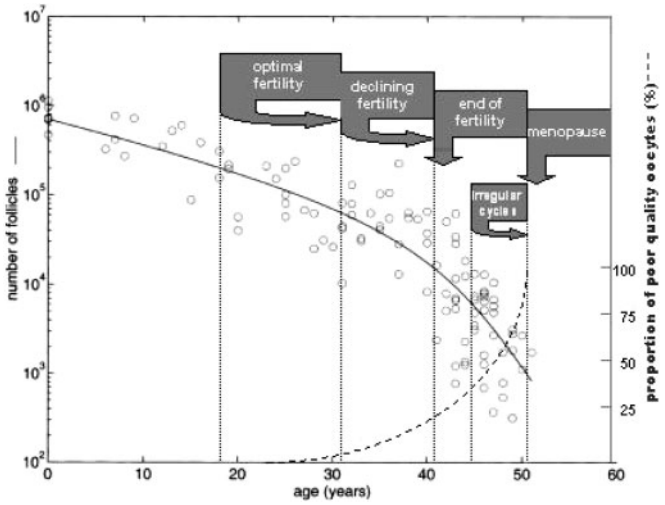

Figure 1. Quantitative (solid line) and qualitative (dotted line) decline of the ovarian follicle pool, which is assumed to dictate the onset of the important reproductive events [reproduced and adapted with permission from de Bruin and te Velde (2004)].

increases from under 5% in women in their early 20s to approximately 30% or over in the age group of 35 years and older (Abma *et al*., 1997). So, although the majority of women of older age will obtain the desired pregnancy within a one-year period, the chance of becoming subfertile increases ~6 fold in comparison with very young women.

The age-related effect on female fertility has also been shown in numerous reports on the results of IVF treatment in infertile couples. The probability of live birth obtained through IVF treatment clearly decreases after the age of 35 (Anonymous, 1995; Templeton *et al*., 1996) and the same has been shown to be true for the implantation rate per embryo (van Kooij *et al*., 1996). In fact, female age has consistently been shown to be an important predictor of success in IVF treatment.

Over the past two decades, a number of so-called ovarian tests have been studied for their ability to predict outcome of IVF in terms of oocyte yield and occurrence of pregnancy. Some of these tests have become part of the routine diagnostic procedure for infertility patients that will undergo assisted reproductive techniques. With the current work we aim to provide an answer to the question of what the true value is of these tests to patient management. Evidence-based medicine has progressively developed as the standard approach for many diagnostic procedures and treatment options in the field of reproductive medicine (National Collaborating Center for Women's and Children's Health, 2004). Therefore, we provide a comprehensive systematic literature review, including an *a priori* protocolized information retrieval on all currently available and applied tests to determine ovarian reserve (OR).

What follows is first a general section in which we briefly outline the aims and the valuation of OR testing and the set-up of the systematic review. After this, we describe individually all currently available tests and their effectiveness with regard to prediction of ovarian response and pregnancy after IVF in generally accepted terms for diagnostic procedures. A unique feature of this systematic review is that we will furthermore provide where possible an integrated receiver operating characteristic (ROC) analysis and curve of all individual evaluated published papers of each test, as well as a formal judgement upon the clinical value.

The assessment of OR

OR can be considered normal in conditions where stimulation with the use of exogenous gonadotrophins will result in the development of at least 8–10 follicles and the retrieval of a corresponding number of healthy oocytes at follicle puncture (Fasouliotis *et al*., 2000). With such a yield, the chances of producing a live birth through IVF are considered optimal. In general, as outlined earlier, age of the woman is a simple way of obtaining information on the extent of her OR, in terms of both quantity and quality (Templeton *et al*., 1996). However, in the view of the substantial variation in the decline of reproductive capacity with age (te Velde and Pearson, 2002) (Figure 2), there is a need to identify women of relatively young age with clearly diminished reserve, as well as women around the mean age at which natural fertility on average is lost (41 years) but still with adequate OR. In clinical terms, we aim to identify women with a high risk of producing a poor response to ovarian stimulation and/or a very low probability of becoming pregnant through IVF, as well as those who still produce enough oocytes to have a good chance of becoming pregnant even if female age is advanced. If it appears possible to identify such categories of women, then management could be individualized, for instance by stimulation dose or treatment scheme adjustments (Tarlatzis *et al*., 2003), by counselling against initiation of IVF treatment or pertinent refusal to accept initiation, or by indicating the necessity of early initiation of treatment before reserve has diminished too far.

OR is currently defined as the number and quality of the follicles left in the ovary at any given time. An accurate measure of the quantitative OR would involve the counting of all follicles present in both ovaries, as is done in post-mortem studies (Block, 1952). For obvious reasons, in OR testing, the true size of the follicle pool has not been used as the benchmark for evaluation (Lass *et al*., 1997a; Lambalk *et al*., 2004; Lass, 2004; Sharara and Scott,

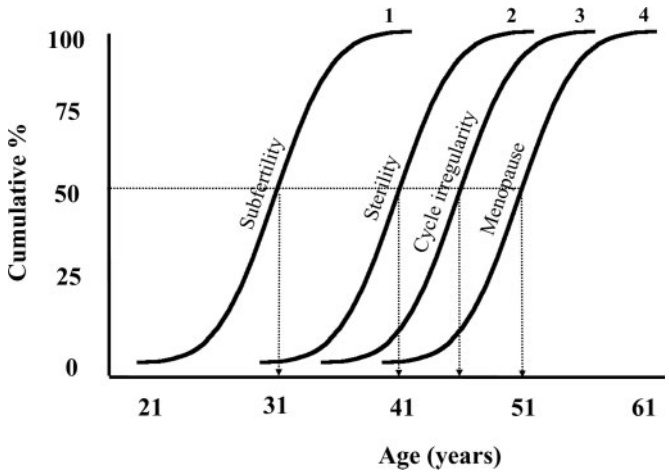

Figure 2. Variations in age at the occurrence of specific stages of ovarian ageing. For explanation of the background of data, see te Velde and Pearson (2002). Reprinted with permission from te Velde and Pearson (2002).

2004), apart from one distinct study (Gulekli *et al*., 1999), where whole ovary counts served as reference for several OR tests (ORTs). Instead, several proxy variables of the pool size are used in studies on diagnostic accuracy, like ovarian response to hyperstimulation with exogenous FSH in IVF and the occurrence of menopause or menopausal transition, as these events are quantitatively determined. Although related, the quality of the oocyte released from the dominant follicle at ovulation represents the other aspect of ovarian reserve. Proxy variables for oocyte quality currently used are the pregnancy probability in infertility treatment like IUI and IVF or in the follow-up of couples during and after the initial infertility work-up.

We should therefore realize that in the vast majority of studies on ORTs that will be discussed below, either ovarian response or occurrence of pregnancy in IVF serves as the benchmark to judge upon the accuracy and clinical value of the test under study. Ovarian response to adequate stimulation may be considered the most accurate, though still indirect, representation of the status of the primordial follicle pool, as it is a condition that is continuously present in the individual that undergoes the test. In contrast, the occurrence of pregnancy in such an individual may be influenced by many more factors than oocyte, and hence embryo quality, alone. Only if the occurrence of pregnancy is studied in a series of treatment cycles it may represent a solid proxy variable of the benchmark for ovarian reserve. Most ORTs are quite adequate in predicting ovarian response, but often fail to correctly predict the occurrence of pregnancy, especially if only one IVF cycle was studied.

Properties of test evaluation

ORT evaluation using response and/or pregnancy as reference or outcome variables should imply the assessment of predictive accuracy and clinical value of the test. Accuracy refers to the degree by which the outcome condition is predicted correctly. Summary statistics of accuracy include *sensitivity* (rate of correct identification of cases with poor response), *specificity* (rate of correct identification of cases without poor response), *likelihood ratio* (LR, how many times more likely particular test results are in patients with poor response than in those without poor response) and *diagnostic odds ratios* (DOR, the odds of positive test results in cases with poor response over the odds of positive test results in those without poor response) (Deeks, 2001; Grimes and Schulz, 2005). To identify all cases that will respond poorly to stimulation without judging many normal responders badly, the test must have high sensitivity and high specificity.

Positive LRs above 10 and negative LRs below 0.1 are considered as indicators of an adequate diagnostic test, while values between 5 and 10 and below 0.2 are considered to indicate a moderate test. As such, the LR can be considered a clinically useful tool to help judge the performance of the test, as the value will change when the threshold for an abnormal test is shifted.

The diagnostic odds ratio is an adequate measure when combining studies in a systematic review, as a single diagnostic odds ratio corresponds to a set of sensitivities and specificities depicted by an ROC curve and is considered threshold independent (Figure 3). It therefore can be considered a good parameter to compare the overall accuracy of a test evaluated in different studies. Although the

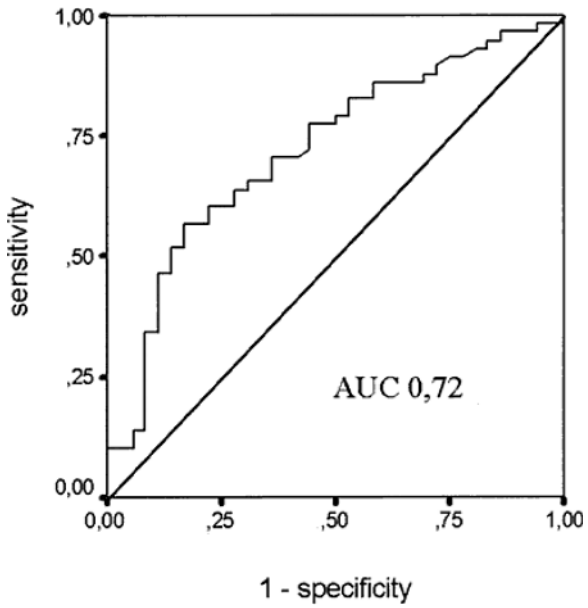

Figure 3. Receiver operator characteristics (ROC) curve depicting the continuous relationship between sensitivity and specificity with shifting threshold values for a given test. The area under the ROC curve (AUC) provides general information on the discriminatory capacity of the test.

DOR values will be higher for tests with better combinations for sensitivity and specificity, this value has not been advocated as a single measure of clinical value, as changes in the threshold used will not be expressed by a change in DOR value. For the meta-analytic approach, the range of DOR values across studies gives some indication as to the homogeneity of such studies.

Finally, the *area under the ROC curve* provides information on the overall discriminatory capacity of the test. Values of 1.0 imply perfect and that of 0.5 indicate completely absent discrimination.

Clinical value incorporates the question whether application of the test at a certain threshold will really change management or costs or safety or success rates on a population basis. It deals with the valuation of false positive and false negative test results in relation to the consequences of these test results for clinical decisions. Also it implies the rate of abnormal test results leading to altered decisions within the population of interest.

Design of ORT studies

Studies on the predictive accuracy and clinical value of ORTs should preferably be prospective in design, should examine cohorts of patients in IVF settings without exclusion of cases with signs of diminished ovarian reserve and patient management should not have been influenced by the test under study (verification bias). Also, evaluation should be equally weighted for every case, thus every case should contribute the same amount of cycles to the analysis. In most studies, only one IVF cycle is studied. A case–control design for the purpose of OR testing bears the disadvantage of retrospection and the absence of a reliable estimate of disease prevalence. The tests under study should in principle be reproducible, both at the laboratory (hormone assays) and at the operator level (ultrasound examination). Also, the outcome of treatment (response and pregnancy), serving as the reference for ovarian reserve, should be clearly defined.

The accuracy in predicting a certain outcome by the test under study should be evaluated by constructing contingency tables at several threshold levels for an abnormal test. Using the calculated sensitivity and specificity from each threshold level, a ROC curve (Figure 3) can be drawn and the calculated area under this curve represents the overall predictive accuracy of the test. Assessment of the clinical value is a complex process in which the applicability in daily practice should become clear. The overall accuracy represented by the ROC curve, the choice of a threshold for abnormality, the rate of abnormal tests at that threshold, the post-test probability of disease (i.e. poor response or non-pregnancy), the valuation of false positive and false negative test results and the consequence for patient management of an abnormal test will all contribute to the process of deciding whether a test is useful or not. Finally, the cost of carrying out the test as a routine measure and the burden to the patient balanced against the reduction in costs by excluding cases with low pregnancy prospects should contribute to the decision whether or not to apply a test.

ORTs in relation to other predictors of success

It is important for patients who are considering treatment with IVF to know the probability of success in the course of a series of IVF treatment cycles. The possibility of a live birth for any couple undergoing treatment will depend on the success rate at the individual clinic. However, equally important in the prediction of outcome are the characteristics of the couple seeking treatment (Stolwijk *et al*., 1996; Templeton *et al*., 1996; Sharma *et al*., 2002). Serious effort has been put into the build-up of prediction models that estimate the probabilities for success prior and during subsequent IVF cycles. In general, these models appeared inaccurate when external validation studies were carried out (Stolwijk *et al*., 1998; Smeenk *et al*., 2000). Intuitively, many IVF centres will use factors like female age, parity, duration of infertility, ovarian response in the first IVF attempt and embryo quality for individual counselling, albeit not through a formal prediction model. Within this practice, ORTs also may play a certain role and female age will be the one ORT applied almost without exception. The pressing question would be to what extent other, endocrine- or ultrasound-based, ORTs contribute and add to the prognostic information already obtained from the infertility work-up or the first IVF cycle. To date, studies specifically addressing this question are scarce or do not include the full range of prognostic factors available.

There are a number of studies (Eimers *et al*., 1994; Collins *et al*., 1995; Snick *et al*., 1997; Hunault *et al*., 2004; Hunault *et al*., 2005) that offer a model, based on factors like duration of subfertility, female age, parity, sperm quality and post-coital test, for the prediction of live birth among untreated subfertile couples. However, none of these models included ORTs, apart from female age. Only one study showed that on top of predictions based on the Eimers model, ORTs failed to add relevant information to the couple's chances for a spontaneous pregnancy (van Rooij *et al*., 2005).

General remarks on physiological background of ORTs

Tests that are used to predict some defined outcome related to ovarian reserve almost without exception give assessment of the number of follicles remaining at some time point in both ovaries. Any marker giving an estimate of the remaining pool will at the same time be capable of providing, to some extent, information on oocyte quality. But on average, from prediction studies it seems that some markers give a better indication of quality than others. Female age, for instance, is the basic factor that is related to both quantity and quality. Basal FSH, through the feedback of inhibin B and estradiol, will represent cohort size but mostly at the extremes and therefore give a more thorough indication of quality aspects. This is in contrast to the more direct quantitative tests using antral follicle count (AFC), anti-Müllerian hormone (AMH) and ovarian volume (OVVOL) that are capable of describing a more complete range of ovarian reserve states. By choosing the right thresholds these tests may eventually correctly predict oocyte quality. The true relation between quantity and quality, however, remains a source of debate. Quantity is an aspect of ovarian reserve that is present in a continuous state and therefore offers a more or less continuous measurability. Quality, however, comes to expression every now and then, even in the setting of IVF. The relationship between the two aspects of ovarian reserve has become more evident when the predictive value of a poor response in a first IVF cycle was examined towards the probability of pregnancy in the actual or subsequent cycles (Klinkert *et al*., 2004). While cases with a normal response in additional cycles yielded acceptable rates of pregnancy, it was shown that in repeated poor responders this probability never surpassed 10% (de Boer *et al*., 2002; Lawson *et al*., 2003; Klinkert *et al*., 2004). It is also important to remember that there are several factors that contribute to the occurrence of pregnancy other than ovarian reserve, such as embryo transfer technique and number of embryos replaced. Even in young women with normal reserve the chance of non-pregnancy remains at least at the 50% level. So, a non-pregnancy state after IVF may even be attributed to unknown, yet non-ovarian reserve related, factors.

Approach of the systematic review

The aim of the systematic review on the value of diagnostic tests is to obtain an overall estimate of the test accuracy and clinical value based on all present evidence, after assessing the quality of the included studies and evaluating the variation in findings among the studies (Irwig *et al*., 1995; Deeks, 2001; Deville *et al*., 2002; Honest and Khan, 2002; Glas *et al*., 2003). Systematic review and meta-analysis on diagnostic accuracy and value implies consecutive steps as summarized in Table I (Irwig *et al*., 1994; Mol *et al*., 1997) please see addendum.

For each study finally included in the meta-analysis, sensitivity and specificity are calculated from the contingency tables. Homogeneity of the sensitivity–specificity points is tested by means of the χ^2-test statistic. A summary point estimate of sensitivity and specificity and the 95% confidence interval is calculated if homogeneity cannot be rejected. In case of heterogeneity, logistic regression is used to evaluate whether Quality/Methodology characteristics of a study are associated with the discriminative capacity of the test under study. If one of the study characteristics is found to have a statistically significant impact on the performance of the test, further analysis is performed in subgroups of patients. If not, it is explored whether the differences in sensitivity–specificity combinations are because of the use of different threshold levels of the test under study. For this purpose, a Spearman correlation coefficient is calculated to assess the association between sensitivity and specificity. If there is a negative correlation as defined by a correlation

Table I. Stepwise approach to the systematic review and meta-analysis of diagnostic tests

1	Define the objective	Test and disease of interest. Reference standard for the disease. Impact of test result on clinical management. Comparison of tests
2	Literature search	Search, link and MESH terms. Inclusion and exclusion criteria. Databases used. Cross references. Contact authors for raw data if appropriate
3	Data extraction	Contingency table. Quality/Methodology characteristics. Extraction by two independent researchers. Disagreement solved by third independent researcher
4	Homogeneity test	Chi-square on sensitivity (sens) and specificity (spec) and provide ROC plot and sens, spec and diagnostic odds ratio (DOR) plot with 95% CI. Focus on outliers
	Homogeneity not rejected	Calculate summary point estimates for sens and spec and 95% CI
	Homogeneity rejected	Logistic regression analysis on relation Quality/Methodology characteristics and test accuracy. If present: subgroup analysis. If absent assume cut-off point effect
5	Data pooling	Spearman correlation between sens and spec ($r < -0.5$) or fixed effect logistic regression of ln DOR with an interaction term for test and study
	Sens and spec related and/or DORs homogenous	Summary ROC curve estimation using random-effects regression model
	Sens and spec not related and/or DORs heterogenous	No pooling possible. Subgroup analysis?
6	Assess clinical value	Positive predictive value of abnormal test at various prevalence values using various thresholds based on summary ROC curve, in correspondence with abnormal test rate If no estimated curve or point: comparison of individual sens and spec points with desired level of sens and spec

coefficient of –0.5 or stronger, the individual pairs of sensitivity and specificity are considered to originate from a single ROC curve. All sensitivity–specificity points are then plotted and a summary ROC curve is estimated using a random-effects regression model (Littenberg and Moses, 1993; Midgette *et al*., 1993; Moses *et al*., 1993).

An important issue is the fact that individual studies may produce highly variable sensitivity–specificity points in the ROC space. This is generally explained by variation in the applied threshold level for an abnormal test across the studies or the presence of considerable study heterogeneity. As in the formal analysis, the presence of heterogeneity in design will be dealt with, and the variation in sens/spec points is generally attributed to the variation in threshold levels and thus allows us to construct a summary ROC curve. At the same time, the threshold variation will prevent the possibility of assessing a single threshold for a specific test that has a generalizable value. This will only become possible if from every study the original database would be available and to date this seems to be an extreme effort.

To assess the clinical value of the test under study for the assessment of disease state (i.e. poor response or non-pregnancy), the positive and negative predictive values are calculated using the estimated summary ROC curve and assuming arbitrary prevalences of the disease in the population. An LR for a positive (or abnormal) test result is then calculated for each point on the estimated ROC curve. Subsequently, the post-test probabilities of disease at various LR values are then calculated for the arbitrary pre-test probabilities of disease, assuming independence between the pre-test probability and the performance of the test (Bancsi *et al*., 2003). Final judgement depends on the overall accuracy, the choice of the test threshold, the post-test prediction at that threshold level and the valuation of a false positive test result. In case no estimated curve from the selected studies can be constructed, the judgement upon the clinical value is based on a comparison of a preset level of sensitivity and specificity with the observed levels in the various studies.

Systematic reviewing of ORTs

The aim of the present series of systematic reviews is to assess the true diagnostic accuracy and clinical value of the ORTs known to date, when applied in an IVF/ICSI population. Reference standards used to valuate the test properties are response to ovarian stimulation and occurrence of pregnancy. No preset definition was used for these standards. For every ORT under study, a computerized MEDLINE search was performed to identify articles on the subject outlined in the previous chapters published until December 2004. Checking of reference lists of articles already obtained was done, all in an iterative fashion. Keywords used for the various searches were ‘in vitro fertilization’ or ‘in vitro fertilisation’ or ‘assisted’ or ‘intracytoplasmatic’ or ‘intracytoplasmic’, in combination with ‘test-specific’ keywords, as mentioned in the tables.

One investigator *(DH or JK)* read all abstracts of the articles that were identified by the search. Any article reporting on the association of the test with poor ovarian response and/or non-pregnancy after IVF or possibly containing information that was to be transformed into a predictive tabulation was pre-selected. Subsequently, all pre-selected articles were fully read and judged independently by two investigators *(DH and JK)*, and separate 2×2 tables were constructed for cross classification of the test result and the occurrence of poor response and/or non-pregnancy, whenever possible. In the event of disagreement on the inclusion or exclusion of pre-selected studies for the meta-analysis or on the calculation of the 2×2 table data or the scoring of quality characteristics, the judgement of a third author *(FB or CL)* was decisive. Studies in which it was not possible to construct 2×2 tables were excluded. Cross-references in all selected articles were checked, and, if applicable, studies were added to the analysis.

Each study was scored by the investigators on the following Quality/Methodology characteristics: (i) sampling (consecutive versus other), (ii) data collection (prospective versus retrospective), (iii) study design (cohort study versus case–control study), (iv) blinding (present or absent), (v) selection bias, (vi) verification bias, (vii) analysis on one or multiple cycles per couple and (viii) definition of outcome, poor response and pregnancy.

In the following sections, the results of search, data extraction, quality and methodology assessment and meta-analysis of extracted data as outlined above are discussed for every ORT comprised in this review.

Basal FSH

Systematic review

Through the search and selection strategy, a total of 37 studies reporting on the capacity of basal FSH to predict poor ovarian response and/or non-pregnancy after IVF and which were suitable for data extraction and meta-analysis were identified (Scott *et al*., 1989; Padilla *et al*., 1990; Toner *et al*., 1991; Khalifa *et al*., 1992; Chan *et al*., 1993; Ebrahim *et al*., 1993; Fanchin *et al*., 1994; Huyser *et al*., 1995; Licciardi *et al*., 1995; Smotrich *et al*., 1995; Balasch *et al*., 1996; Csemiczky *et al*., 1996; Martin *et al*., 1996; Pruksananonda *et al*., 1996; Gurgan *et al*., 1997; Chang *et al*., 1998a; Evers *et al*., 1998; Ranieri *et al*., 1998; Sharif *et al*., 1998; Bassil *et al*., 1999; Hall *et al*., 1999; Bancsi *et al*., 2000; Chae *et al*., 2000; Creus *et al*., 2000; Fabregues *et al*., 2000; Jinno *et al*., 2000; Penarrubia *et al*., 2000; Mikkelsen *et al*., 2001; Nahum *et al*., 2001; van der Stege and van der Linden, 2001; Esposito *e al*., 2002; Chuang *et al*., 2003; Fiçicioğlu *et al*., 2003; Kwee *et al*., 2003; Yanushpolsky *et al*., 2003; Akande *et al*., 2004; Erdem *et al*., 2004). Characteristics of the included studies are listed in Table II. As shown, there was a large diversity with regard to the various aspects of methodology and quality, and the definition of poor ovarian response. Logistic regression analysis indicated no significant association between any of these study characteristics and the predictive performance of basal FSH. For example, whether the design of the study was retrospective or prospective did not influence the prognostic capacity of basal FSH.

Accuracy of poor response prediction

The sensitivities and specificities, as well as the positive LRs of an abnormal test and the DORs for the prediction of poor ovarian response, as calculated from each study, are summarized in Table III, please see addendum. Sensitivity and specificity points, as plotted in Figure 4, were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity 0.001 and *P*-value for specificity 0.001). Therefore, calculation of one summary point estimate for sensitivity and specificity was not meaningful for overall judgement of accuracy. The Spearman correlation coefficient for sensitivity and specificity was –0.87, which was judged to be sufficient to estimate a summary ROC curve (Figure 4).

Accuracy of non-pregnancy prediction

Sensitivities and specificities for the prediction of non-pregnancy, as calculated from each study, are summarized in Table IV, please see addendum. Again, sensitivity and specificity points plotted in Figure 5 were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity 0.001 and *P*-value for specificity 0.001). The Spearman correlation coefficient for sensitivity and specificity was –0.82 and as such was sufficient to estimate a summary ROC curve (Figure 5).

Clinical value

Based on the summary ROC curves depicted in Figure 4, a range of positive LRs was calculated and for each ratio the pre-FSH test probability of poor response and non-pregnancy was converted into a post-FSH-test probability. Table V, (please see addendum) depicts the probability of obtaining a certain FSH test result and the corresponding LR within different LR ranges for the prediction of poor response and non-pregnancy. At a maximum positive LR of 8, the post-FSH-test probability of poor response will approximate 70% if the pre-FSH-test probability is assumed to be as high as 20%. As is apparent from this table, the probability of obtaining a test result (FSH level) with an LR of ~8 is quite small. Table III shows that in women with an increased FSH level the probability of poor response only increases substantially (3-fold or more) in studies applying a high threshold level for FSH, resulting in a very limited number of patients with an abnormal test result.

Even more so, for prediction of non-pregnancy, the extremely high FSH levels that are necessary to obtain the moderate positive LR of ~5, leading to a post-test pregnancy rate of less than 5%

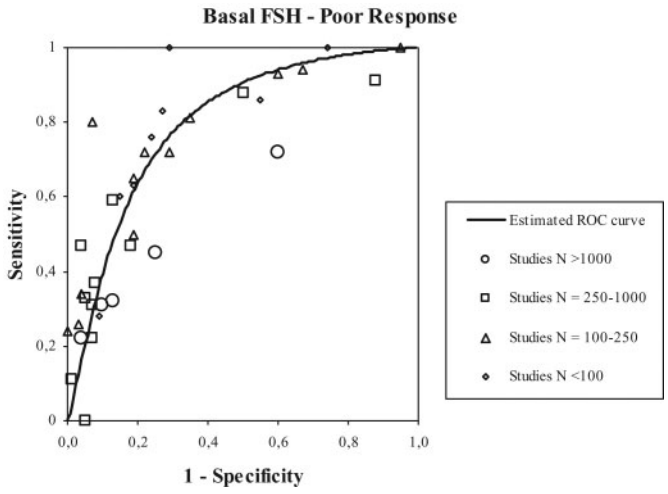

Figure 4. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of basal FSH in the prediction of poor response. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

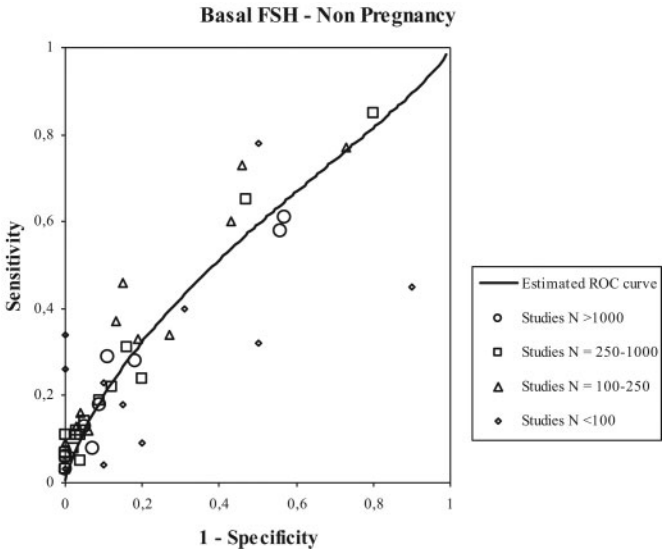

Figure 5. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of basal FSH in the prediction of non-pregnancy. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

based on a pre-test rate of 20%, again occur only in a very limited number of patients (Table V). Beyond the coordinate defined by specificity 0.90 and sensitivity 0.20, the summary ROC curve almost runs parallel to the line of equality. This indicates that this segment of the curve is 100% uninformative (LR ~1).

All this leads to the conclusion that with the use of basal FSH in regularly cycling women, accuracy in the prediction of poor response and non-pregnancy is adequate only at very high threshold levels, but because of the very low numbers of abnormal tests has hardly any clinical value. Considering this along with a false positive rate of ~ 5%, the test will not be suitable as a diagnostic test to exclude patients, but only as screening test for counselling purposes and further diagnostic steps, in which a first IVF attempt may be the step of choice (Roberts *et al*., 2005).

AMH

Systematic review

Through the search and selection strategy, two studies reporting on the predictive capacity of AMH and which were suitable for data extraction and meta-analysis were identified (van Rooij *et al*., 2002; Muttukrishna *et al*., 2004). Characteristics of the included studies are listed in addendum, Table VI.

Accuracy of poor response prediction

The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response, as calculated from each study, are summarized in Table VII, (see addendum) and in Figure 6. Homogeneity could not be rejected for sensitivity and specificity (χ^2-test statistic: *P*-value for sensitivity 0.12 and *P*-value for specificity 0.64), but this is merely because of the fact that only two studies were included. As can be seen from Figure 6, the points of the two studies can be thought of as originating from a single ROC curve (Spearman correlation coefficient between sensitivity and specificity is –0.81). The summary ROC curve that can be estimated from these points is also shown in Figure 6.

Accuracy of non-pregnancy prediction

Sensitivities and specificities for the prediction of non-pregnancy by AMH, as calculated from each study, are summarized in Table VIII. As the study of Van Rooij was the only one detected, further meta-analysis is not useful. The ROC-curve derived from the data of Van Rooij *et al*. representing the accuracy of AMH in the prediction of non-pregnancy is shown in Figure 7.

Clinical value

As data from only two studies are available, it is not feasible to extract data on the interrelation between positive LRs, post-test probabilities and the rate of abnormal tests. However, looking at the performance of AMH in the prediction of poor response, a desired level for sensitivity of 75% and for specificity of 85% would imply that the test performs only moderately, especially at the sensitivity level. For non-pregnancy prediction, a desired level of sensitivity of 40% and specificity of 95% would imply that the test has hardly any value, unless very low threshold levels would be used, which will certainly lead to only very small percentages of abnormal tests. Additional studies are to be awaited to learn whether test capacity may prove to be more superior than current tests like basal FSH and the AFC (Hazout *et al*., 2004; Muttukrishna *et al*., 2005; Penarrubia *et al*., 2005).

Inhibin B

Systematic review

We detected a total of nine studies reporting on the predictive capacity of inhibin-B and which were suitable for data extraction and meta-analysis (Balasch *et al*., 1996; Seifer *et al*., 1997; Hall *et al*., 1999; Creus *et al*., 2000; Fabregues *et al*., 2000; Penarrubia *et al*., 2000; Bancsi *et al*., 2002a; Fiçicioğlu *et al*., 2003; Erdem *et al*., 2004). Characteristics of the included studies are listed in addendum Table IX. Variation among the definitions of

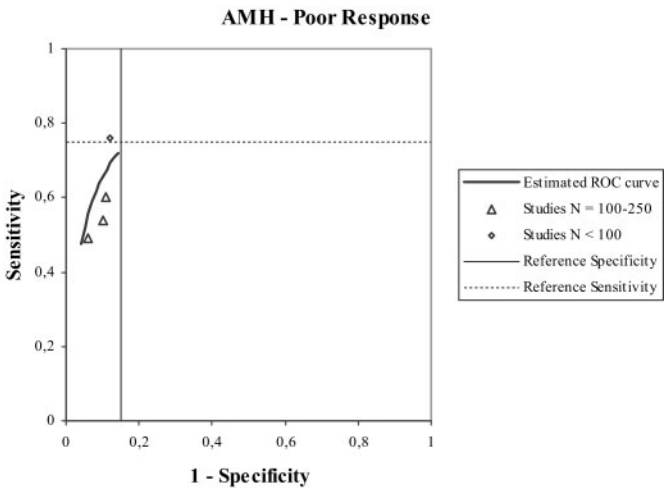

Figure 6. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of AMH in the prediction of poor response. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated. Reference lines indicate a desired level for sensitivity (0.75) and specificity (0.85).

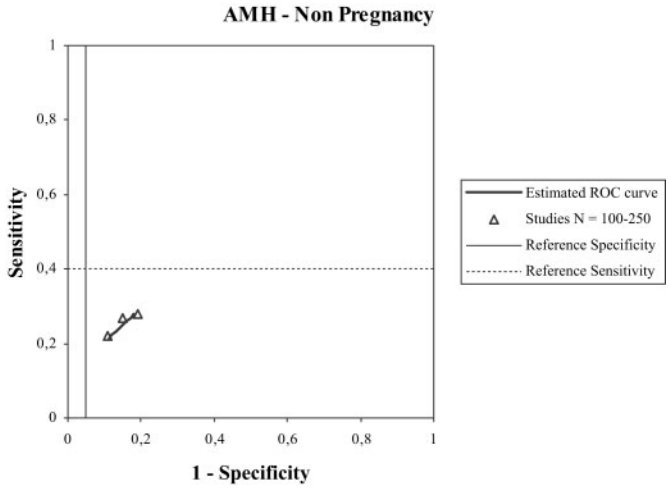

Figure 7. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of AMH in the prediction of non-pregnancy. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated. Reference lines indicate a desired level for sensitivity (0.75) and specificity (0.85).

poor response and study quality and design characteristics was clearly present but logistic regression analysis revealed that none of the items significantly impacted upon the predictive performance of the test. Subgroup analysis therefore was not indicated.

Accuracy of poor response prediction

The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response, as calculated from each study, are summarized in Table X, see addendum. Calculation of one summary point estimate for sensitivity and specificity was not meaningful, as both test characteristics, as plotted in Figure 8, were heterogeneous among studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity 0.002). The Spearman correlation coefficient for sensitivity and specificity was sufficient to estimate a summary ROC curve ($R = -0.93$, Figure 8). In the figure, it is clearly seen that all but one study were close to the estimated ROC curve, and that one study reported a clearly better accuracy (Fiçicioğlu *et al.*, 2003). This study was of good quality, but reported on only a small number of patients.

Accuracy of non-pregnancy prediction

There were three studies that reported on the capacity of inhibin B to predict non-pregnancy. Sensitivities and specificities for the prediction of non-pregnancy, as calculated from each study, are summarized in Table XI. Sensitivity and specificity as plotted in Figure 9 were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity 0.004 and *P*-value for specificity <0.001). The Spearman correlation between sensitivity and specificity showed a coefficient of –0.94, sufficient to estimate a summary ROC curve.

Clinical value

Based on the summary ROC curves depicted in Figure 8, a range of positive LRs was calculated and for each ratio pre-inhibin B-test probabilities of poor response or non-pregnancy (20 and 80%, respectively) were converted into post-inhibin B-test probabilities. Table XII depicts the probability of obtaining a certain inhibin B test result and the corresponding LR, within different LR ranges for the prediction of poor response and non-pregnancy. At a very modest LR of 4, the post-inhibin B-test probability of poor response will not be higher than 55%, while the chance of obtaining such a test result is very small.

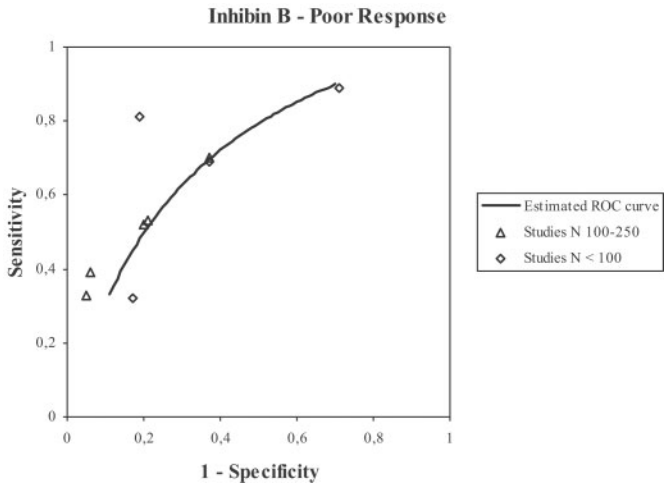

Figure 8. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of inhibin B in the prediction of poor response. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

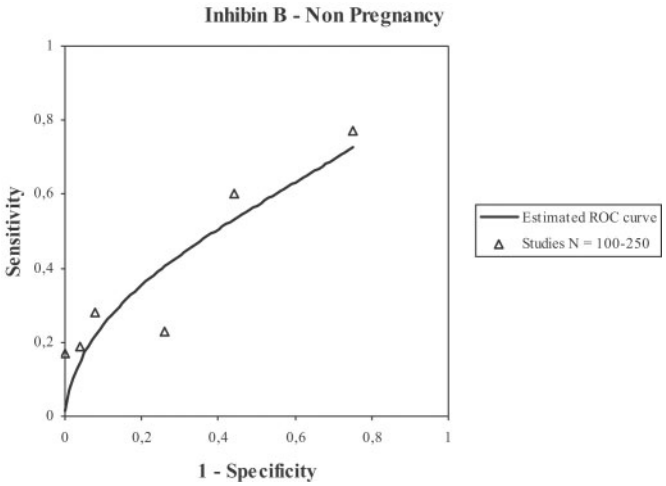

Figure 9. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of inhibin B in the prediction of non-pregnancy. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

For prediction of non-pregnancy, extreme threshold levels are necessary to obtain a modest positive likelihood ratio of ~4–5, leading to a post-test pregnancy rate of approximately 5%. Such abnormal test results occur only in a very limited number of patients, while the false positive rate will lead to unnecessary exclusions from IVF programs if the test is used in a diagnostic fashion.

With the use of basal inhibin B in regularly cycling women, the accuracy in the prediction of poor response and non-pregnancy is only modest at a very low threshold level. At best the test may be used as screening test for counselling purposes or to direct further diagnostic steps, like a first IVF attempt to observe the response to ovarian stimulation. Used in this way, the test may well be inferior to other tests discussed in this review.

Basal estradiol

Systematic review

We detected a total of 10 studies reporting on the predictive capacity of basal estradiol and which were suitable for data extraction and meta-analysis (Licciardi *et al.*, 1995; Smotrich *et al.*, 1995; Evers *et al.*, 1998; Vazquez *et al.*, 1998; Hall *et al.*, 1999; Frattarelli *et al.*, 2000; Penarrubia *et al.*, 2000; Phophong *et al.*, 2000; Mikkelsen *et al.*, 2001; Ranieri *et al.*, 2001; Bancsi *et al.*, 2002a). Characteristics of the included studies are listed in addendum Table XIII. Again, variation among the definitions of poor response and study quality and design characteristics was clearly present, but logistic regression analysis revealed that none of the items significantly impacted upon the predictive performance of the test. Subgroup analysis therefore was not indicated.

Accuracy of poor response prediction

There were eight studies that reported on the prediction of poor response. The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response, as calculated from each study, are summarized in Table XIV. Calculation of one summary point estimate for sensitivity and specificity was not meaningful, as both test characteristics as plotted in Figure 10 were heterogeneous among studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity 0.002). The Spearman correlation coefficient for sensitivity and specificity was –0.50. As can be seen from Figure 10, this can be because of three outliers, which were extracted from the studies of Smotrich *et al*. and Ranieri *et al*. From neither the clinical nor the methodological point of view could a clear explanation be provided for the outliers. When correlation between sensitivity and specificity was assessed after exclusion of the three outliers, we found a very strong correlation (–0.94). Figure 10 shows two estimates of a summary ROC curve, one constructed with all data and one constructed after exclusion of the two studies with outlying data (Figure 10).

Accuracy of non-pregnancy prediction

There were nine studies that reported on the capacity of basal estradiol to predict non-pregnancy after IVF. Sensitivities and specificities for the prediction of non-pregnancy, as calculated from each study, are summarized in Table XV. Again, sensitivity and specificity as plotted in Figure 11 were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity <0.001). The Spearman correlation between sensitivity and specificity showed a coefficient of –0.89, sufficient to estimate a summary ROC curve (Figure 11). This summary ROC curve is almost parallel to the line $x = y$, indicating virtually no discriminative capacity.

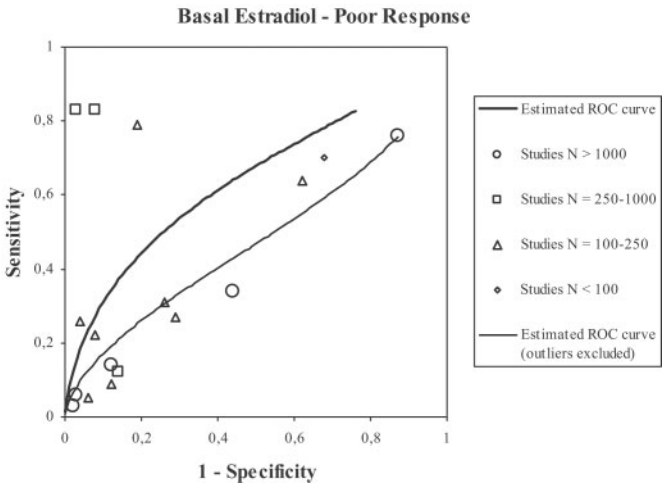

Figure 10. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of basal estradiol in the prediction of poor response. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

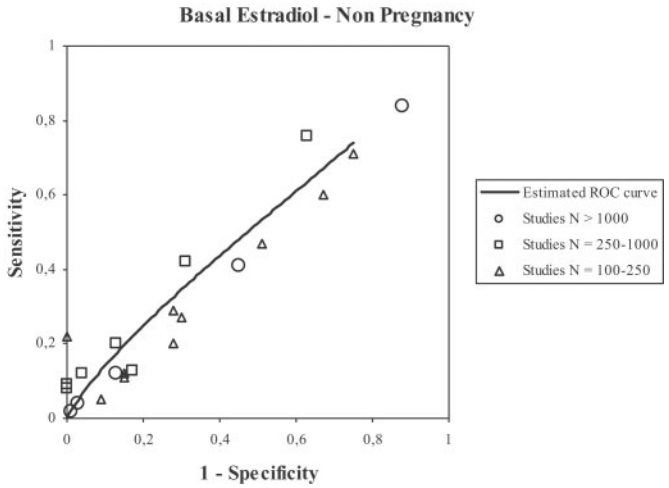

Figure 11. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of basal estradiol in the prediction of non-pregnancy. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

Clinical value

Based on the two summary ROC curves for all studies depicted in Figure 10, a range of positive LRs was calculated and for each ratio, pre-estradiol-test probabilities of poor response or non-pregnancy (20 and 80%, respectively) were converted into post-estradiol-test probabilities. Table XVI (please see addendum) depicts the probability of obtaining a certain estradiol-test result and the corresponding LR, within different LR ranges for the prediction of poor response and non-pregnancy. At a moderate LR of 4–5, the post-estradiol-test probability of poor response will not be higher than ~50%, while the chance of obtaining such a test result is very small.

For prediction of non-pregnancy no clear threshold levels can be identified for basal estradiol that will lead to an adequate combination of LR, post-test probability and abnormal test rate. This could be anticipated from the shape of the ROC curve in Figure 11

All this leads to the conclusion that the clinical applicability for basal estradiol as a test before starting IVF is prevented by the very low predictive accuracy, both for poor response and non-pregnancy.

AFC

Systematic review

Through the search and selection strategy, a total of 15 studies reporting on the predictive capacity of basal AFC and suitable for data extraction and meta-analysis were identified (Chang *et al*., 1998b; Frattarelli *et al*., 2000; Ng *et al*., 2000; Sharara and McClamrock, 2000; Hsieh *et al*., 2001; Nahum *et al*., 2001; Bancsi *et al*., 2002a; Erdem *et al*., 2002; Fisch and Sher, 2002; Fiçicioğlu *et al*., 2003; Frattarelli *et al*., 2003; Jarvela *et al*., 2003; Kupesic *et al*., 2003; Yong *et al*., 2003; Durmusoglu *et al*., 2004). Characteristics of the included studies are listed in addendum Table XVII. Variation among the definitions of poor response and study quality and design characteristics is clearly present but logistic regression analysis revealed that none of the items significantly

impacted upon the predictive performance of the test. Subgroup analysis therefore was not indicated.

Accuracy of poor response prediction

The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response, as calculated from each study, are summarized in Table XVIII. Calculation of one summary point estimate for sensitivity and specificity was not meaningful, as both test characteristics as plotted in Figure 12 were heterogeneous among studies (χ^2-test statistic: *P*-value for sensitivity 0.001 and *P*-value for specificity 0.001). The Spearman correlation coefficient for sensitivity and specificity was –0.57 and was judged to be sufficient to estimate a summary ROC curve (Figure 12).

Accuracy of non-pregnancy prediction

Sensitivities and specificities for the prediction of non-pregnancy, as calculated from each study, are summarized in Table XIX. Again, sensitivity and specificity as plotted in Figure 13 were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity 0.001 and *P*-value for specificity 0.001). The Spearman correlation between sensitivity and specificity showed a coefficient of –0.66, sufficient to estimate a summary ROC curve (Figure 13).

Clinical value

Based on the summary ROC curves depicted in Figure 12, a range of positive LRs was calculated and for each ratio pre-AFC test probabilities of poor response or non-pregnancy were converted into a post-AFC-test probability. Table XX depicts the probability of obtaining a certain AFC test result and the corresponding LR within different LR ranges for the prediction of poor response and non-pregnancy. At a maximum positive LR of ~8, the post-AFC test probability of poor response will approximate 70%, if the pre-AFC-test probability is assumed to be as high as 20%. The probability of obtaining a test result (AFC) with a likelihood ratio ~8 is high enough to consider the AFC as a clinically valuable test for poor response prediction.

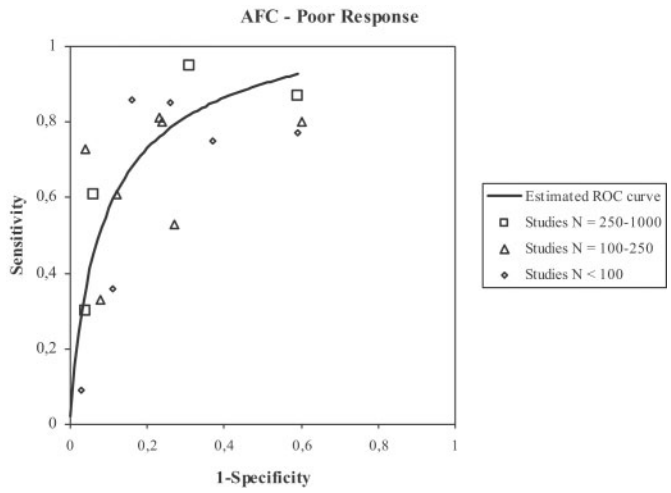

Figure 12. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of the AFC in the prediction of poor response. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

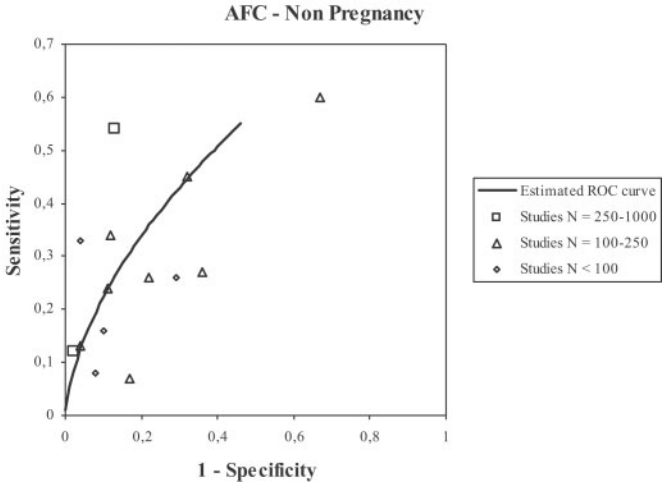

Figure 13. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of the AFC in the prediction of non-pregnancy. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

For prediction of non-pregnancy, the extremely low AFC that is necessary to obtain a moderate positive likelihood ratio of ~5, leading to a post-test pregnancy rate of less than 5% based on a pre-test rate of 20%, occurs only in an extremely limited number of patients (Table XX). Beyond the coordinate defined by specificity 0.80 and sensitivity 0.30, the summary ROC curve almost runs parallel to the line of equality. This indicates that this segment of the curve is 100% uninformative (LR ~1).

Based on these data, it can be concluded that the accuracy of the AFC for predicting poor response in regularly cycling women is adequate at a low threshold level, but because of the very limited numbers of abnormal tests has hardly any clinical value for pregnancy prediction. Added to the false positive rate of ~5% the test will not be suitable as diagnostic test to exclude patients on the basis of the presumed diagnosis of advanced ovarian ageing. It may well be used as a screening test for possible poor responders and for directing further diagnostic steps like a first IVF attempt, where the ovarian response to hyperstimulation will provide additional information (Hendriks *et al.*, 2005d).

OVVOL

Systematic review

For assessing the predictive value of OVVOL, the search detected a total of 10 studies available for data extraction and meta-analysis. Of these, two studies reported solely on the prediction of poor response (Sharara and McClamrock, 1999; Fiçicioğlu *et al.*, 2003) and eight studies reported on the prediction of both poor response and pregnancy (Syrop *et al.*, 1995; Lass *et al.*, 1997b; Frattarelli *et al.*, 2000; Schild *et al.*, 2001; Bancsi *et al.*, 2002a; Jarvela *et al.*, 2003; Kupesic *et al.*, 2003; Erdem *et al.*, 2004). Study characteristics of the included studies are listed in addendum Table XXI. Selection bias was present in almost half of all studies (Lass *et al.*, 1997b; Frattarelli *et al.*, 2000; Kupesic *et al.*, 2003; Erdem *et al.*, 2004). In three studies, patients were selected by basal FSH level (Frattarelli *et al.*, 2000; Kupesic *et al.*, 2003; Erdem *et al.*, 2004) and in the study by Lass *et al.* (Lass *et al.*,

1997b) only patients aged >36 years with an FSH level <15 IU/L were included. Three studies showed evidence of verification bias (Jarvela *et al*., 2003; Kupesic *et al*., 2003; Erdem *et al*., 2004), implying that smaller OVVOL altered the management of the patient by applying higher FSH dosages.

Accuracy of poor response prediction

Sensitivities and specificities, positive LR and the DOR for the prediction of poor ovarian response are summarized in Table XXII. Homogeneity for both sensitivity and specificity had to be rejected (χ^2-test: both *P*-values <0.001). Hence, the calculation of a summary point estimate for sensitivity and specificity was not meaningful. None of the study characteristics recorded had a statistically significant impact on the reported predictive performance of OVVOL. The Spearman correlation coefficient for the relation between sensitivity and specificity was –0.55, sufficient to estimate a summary ROC curve. This curve showed a modest overall predictive accuracy as can be seen in the ROC space in Figure 14.

Accuracy of non-pregnancy prediction

For the prediction of non-pregnancy, test characteristics for each study are summarized in Table XXIII. As with the data for ovarian response, homogeneity for sensitivity had to be rejected. However, specificity appeared to be homogeneous (χ^2-test: *P*-value 0.11). Because for the estimation of one summary point for sensitivity and specificity statistical homogeneity, both test parameters are required, this solution was abandoned. Logistic regression analysis showed that three studies which suffered from verification bias reported a significantly different accuracy compared to the seven remaining studies (p-value: 0.01). None of the other study characteristics had a significant impact on the estimates of test accuracy. In the subgroup analysis of the seven studies without verification bias, homogeneity was again rejected for sensitivity, while specificity again showed homogeneity. The Spearman correlation coefficient for sensitivity and specificity was –0.94, which was judged to be sufficient to estimate a summary ROC curve.

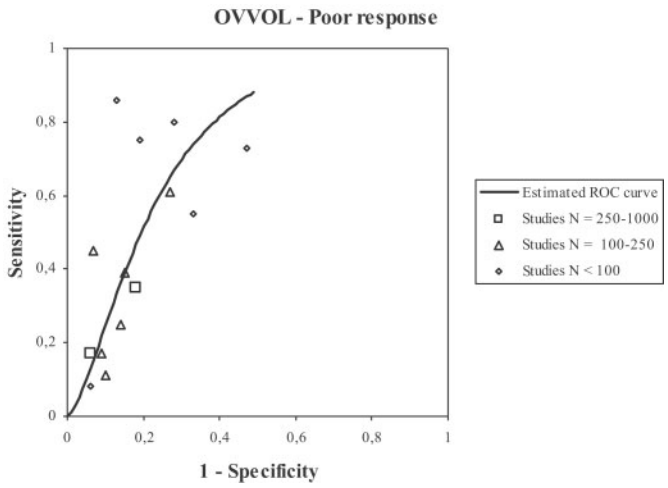

Figure 14. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of OVVOL (ovarian volume) in the prediction of poor response. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

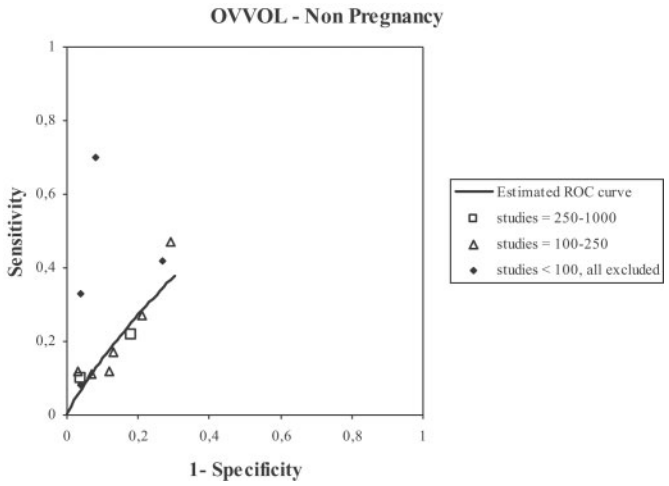

Figure 15. Estimated ROC curve and sensitivity–specificity points for studies reporting on the performance of OVVOL (ovarian volume) in the prediction of non-pregnancy, after exclusion of three studies with verification bias. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

The curve in Figure 15 indicates that OVVOL volume has no clear accuracy in the prediction of non-pregnancy in IVF patients, even if a very low threshold for abnormality of the test would be chosen.

Clinical value

Based on the estimated ROC curves in Figure 14 the probability of obtaining a certain test result for the OVVOL measurement is shown in Table XXIV within a corresponding range of LRs for the prediction of poor ovarian response. Only at modest LRs, the post-test probability of poor response may approach 50%, while abnormal test results will be obtained in some 30% of tested cases. However, applying a more adequate positive likelihood level will result in virtually no cases being identified by the test. For non-pregnancy prediction, Table XXIV shows that for higher LRs (>4), the post-test probability of non-pregnancy may increase to ~93–97%, assuming a pre-test probability of 80%. However, the probability that a test result will be in that range is close to zero. As false positive test results for both ovarian response and non-pregnancy prediction are not acceptable if patients are refused treatment, all this implies that the OVVOL is hardly suitable as a routine test for ovarian reserve assessment.

Ovarian vascular flow

Systematic review

Through the search we detected seven studies reporting on the predictive capacity of ovarian vascular flow parameters for ovarian response and/or the occurrence of pregnancy (Zaidi *et al*., 1996; Engmann *et al*., 1999a,b; Kim *et al*., 2002; Kupesic and Kurjak, 2002; Kupesic *et al*., 2003; Popovic-Todorovic *et al*., 2003b). In these studies ovarian flow was assessed either on cycle day 3 or after achievement of pituitary suppression with a GnRh agonist and before the onset of ovarian stimulation. As only the 2003 study by Kupesic (Kupesic *et al*., 2003) could be included on a 2×2

for cross classification of the test result and the occurrence of poor response or non-pregnancy, it was not possible to carry out a formal meta-analysis (see addendum Table XXV and XXVI). Also, the studies used very different flow-derived predictors. Peak systolic velocity was used as the main predictor (Kupesic *et al*., 2003). Others used ovarian stromal blood flow obtained by 3D power Doppler (Engmann *et al*., 1999a).

Ovarian biopsy

Ovarian reserve depends on the number of primordial follicles in the ovarian cortex, which suggests that the obvious way to obtain an estimate would be to measure follicular density in an ovarian biopsy (Lass, 2001; Lass, 2004). Attempts were made to quantify the number of small antral follicles in small shallow biopsies taken during diagnostic laparoscopy from infertility patients (Lass *et al*., 1997a) and there was a clear age-dependent decline in follicular density. Women over 35 years of age had only 30% of the quantities present in younger women. The number of follicles per unit of volume found in the biopsies was used to estimate the total and it was suggested that it could as such be potentially applied at the individual level. It was recognized though that the biopsy follicle density would not accurately represent the density in the whole ovary (Lass, 2001) and this seems indeed the case. Recently, several investigators have shown that follicle density varied greatly in small pieces of cortex, rendering information from biopsies as completely unreliable for an individual ovarian follicle content irrespective of how many were taken, their size and the location (Qu *et al*., 2000; Schmidt *et al*., 2003; Lambalk *et al*., 2004; Sharara and Scott, 2004). This indicates that the technique which is invasive and potentially harmful in terms of risks of adhesions and other complications of the surgical procedure is intrinsically unreliable and should therefore not be used to evaluate individual ovarian reserve. It is probably useful for research purposes to determine follicle density statistics in patient groups provided that group sizes are such that they compensate for the inherent extreme inter-biopsy and inter-individual spread of information (Qu *et al*., 2000; Schmidt *et al*., 2003; Webber *et al*., 2003; Lambalk *et al*., 2004). Finally, in the context of the current systematic review, there are no studies published that have evaluated ovarian biopsy follicle density for prediction of IVF outcome in terms of ovarian response and pregnancy rates.

Clomiphene Citrate Challenge Test

Systematic review

The computerized MEDLINE search detected 12 studies on the capacity of the Clomiphene Citrate Challenge Test (CCCT) to predict poor ovarian response and/or pregnancy after IVF (Tanbo *et al*., 1989; Loumaye *et al*., 1990; Tanbo *et al*., 1990; Tanbo *et al*., 1992; Csemiczky *et al*., 1996; Kahraman *et al*., 1997; van der Stege and van der Linden, 2001; Csemiczky *et al*., 2002; Kwee *et al*., 2003; Yanushpolsky *et al*., 2003; Erdem *et al*., 2004; Hendriks *et al*., 2005a). Study characteristics of the included studies are listed in addendum Table XXVII. This table shows that many studies suffered from various sources of potential bias, especially selection bias. Also, definitions applied for poor ovarian response and for an abnormal CCCT result (based on either day-10 FSH alone or on both basal FSH and day-10 FSH results) varied considerably. Logistic regression analysis indicated that none of the study characteristics had a statistically significant impact on the reported predictive performance of the CCCT, neither for the outcome response nor for the outcome non-pregnancy. As a consequence, all studies were taken together for further analysis.

Accuracy of poor response prediction

For the prediction of ovarian response, sensitivities and specificities of each study are summarized in Table XXVIII. Homogeneity could not be rejected for sensitivity (χ^2-test statistic: *P*-value 0.09), but had to be rejected for specificity (χ^2-test statistic: *P*-value <0.001). Therefore, calculation of one summary point estimate for sensitivity and specificity was not feasible. Moreover, values of the DOR (range 2.4–38.8) from the various studies appeared heterogeneous, indicating that the individual ROC curves were quite heterogeneous. Also, the Spearman correlation coefficient for sensitivity and specificity values was –0.46, which was judged not to be sufficient to estimate a summary ROC-curve. A plot of the sensitivity–specificity points in an ROC space is shown in Figure 16, showing the considerable heterogeneity which appeared not be attributable to differences in threshold level used.

Accuracy of non-pregnancy prediction

For the prediction on non-pregnancy, the sensitivities and specificities of each study are summarized in Table XXIX. Homogeneity was rejected for both sensitivity and specificity (χ^2-test statistic: *P*-value <0.001 and 0.04, respectively) and calculation of one summary point estimate for sensitivity and specificity was not meaningful. Also, the values of the DOR in the various studies (range 1.0–35.4) appeared non-homogeneous. A plot of sensitivity–specificity points in an ROC space is shown in Figure 17. The Spearman correlation between sensitivity and specificity was –0.20, which again was judged not to be sufficient to estimate a summary ROC curve.

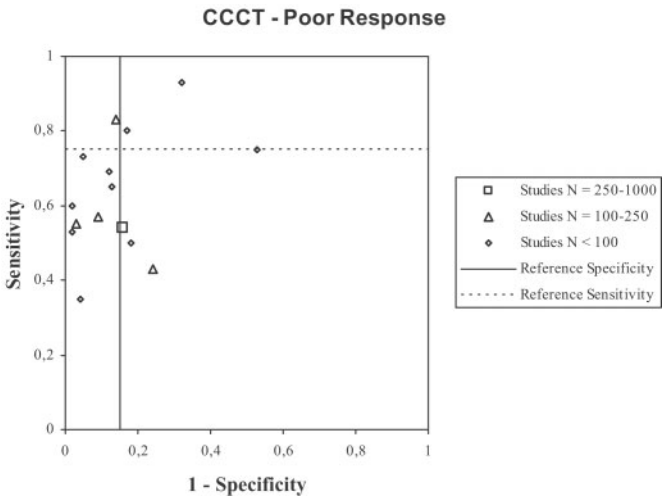

Figure 16. Sensitivity–specificity points for all studies reporting on the performance of the CCCT in the prediction of poor response. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. Because of heterogeneity among studies, no estimated summary ROC point of curve could be constructed. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated. Reference lines indicate a desired level for sensitivity (0.75) and specificity (0.85).

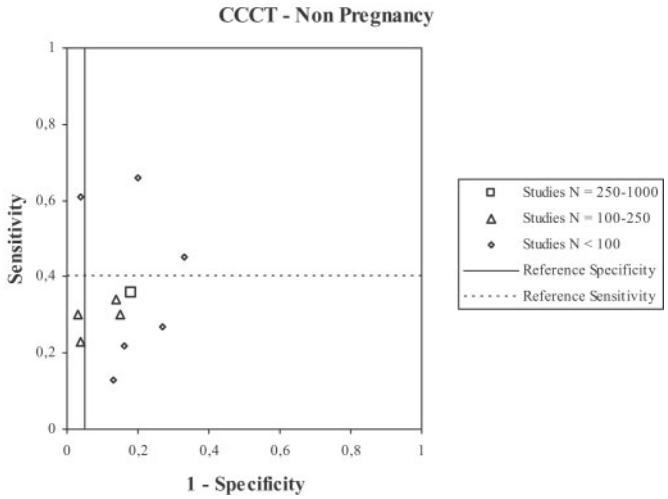

Figure 17. Sensitivity–specificity points for all studies reporting on the performance of the CCCT in the prediction of non-pregnancy. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. Because of heterogeneity among studies no estimated summary ROC point of curve could be constructed. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated. Reference lines indicate a desired level for sensitivity (0.75) and specificity (0.85)

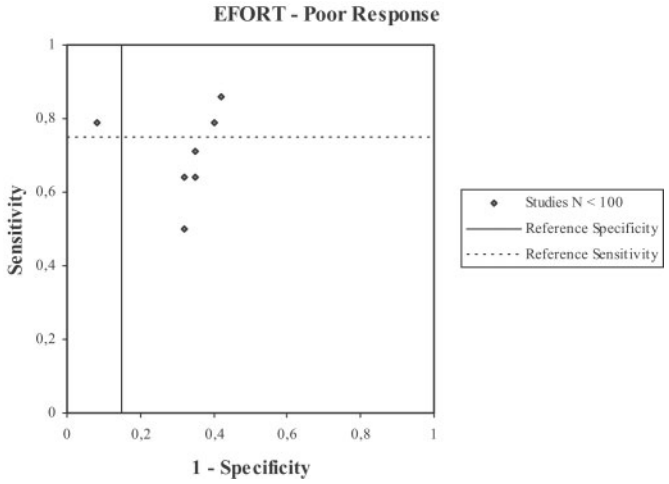

Figure 18. Sensitivity–specificity points for all studies reporting on the performance of the EFORT in the prediction of poor response. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. Because of heterogeneity among studies no estimated summary ROC point of curve could be constructed. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated. Reference lines indicate a desired level for sensitivity (0.75) and specificity (0.85)

Clinical value

Because of the absence of estimated ROC curves for response and non-pregnancy prediction, the interrelation between positive LR, post-test probability and percentage of abnormal tests could not be calculated. It is considered that a challenge test used as a diagnostic tool to identify poor responders should have sensitivity and specificity at a certain desired level. If these levels are set at 75 and 85%, respectively, it can be concluded from Figure 16 that hardly any study will fulfil these criteria. Moreover, in comparative studies the clinical performance of the CCCT in response prediction appeared not better than that of the AFC or FSH (Jain *et al*., 2004; Hendriks *et al*., 2005c). Regarding prediction of non-pregnancy, desired levels for a test that excludes cases from entering an IVF program should arbitrarily be set at 40% for sensitivity and 95% for specificity. The vast majority of studies fail to reach both criteria as shown in Figure 17 As such the CCCT performs no better than other tests like the AFC or basal FSH, especially because of a loss in specificity.

Exogenous FSH ORT

Systematic review

We detected three studies from the literature reporting on the predictive capacity of the exogenous FSH ORT (EFORT) that were suitable for data extraction (Fanchin *et al*., 1994; Kwee *et al*., 2003; Yong *et al*., 2003). The characteristics of these studies are listed in addendum Table XXX.

Accuracy of poor response prediction

The individual values for sensitivity and specificity pairs are summarized in Table XXXI and plotted in Figure 18. As can be seen from this ROC space, the three detected studies report sensitivities around 80%, whereas specificities vary around 60% in the study of Kwee *et al.* and Yong *et al.* and above 90% in the study of Fanchin *et al*. In view of these different results between the studies, further assessment of heterogeneity appeared not useful and therefore a summary point or curve in the ROC space could not be constructed.

Accuracy of non-pregnancy prediction

No single study reported on the predictive accuracy using the outcome pregnancy as test reference.

Clinical value

Because of the absence of an estimated ROC curve for poor response prediction, the interrelation between positive LR, post-test probability and percentage of abnormal tests could not be calculated. It is considered that a challenge test used as a diagnostic tool to identify poor responders should have sensitivity and specificity at a certain desired level. If these levels are set at a minimum level of 75 and 85%, respectively, it can be concluded from Figure 18 that only one study fulfils these criteria (Fanchin *et al*., 1994). Especially, the false positive prediction may hamper the use of this test if a high level of detection is needed and patients are refused IVF on the basis of the test result. Finally, in comparison to basal tests, challenge tests should clearly improve prediction if they are to be preferred.

Gonadotrophin: releasing hormone agonist stimulation test

Systematic review

Through the search and selection strategy, a total of four studies reporting on the predictive capacity of the Gonadotrophin releasing hormone agonist stimulation test (GAST) were identified and considered suitable for data extraction and meta-analysis (Ranieri *et al*., 1998; Padilla *et al*., 1990; Winslow *et al*., 1991; Hendriks *et al*., 2005b). Characteristics of the included studies are listed in addendum Table XXXII. Considerable variation among the definitions of poor response and the study quality and design characteristics was observed, but as only three studies reported on each of

the two endpoints, a systematic analysis of these study characteristics was not indicated.

Accuracy of poor response prediction

There were three studies that reported on the prediction of poor response. The sensitivities and specificities, the positive LR and the DOR for the prediction of poor ovarian response, as calculated from each study, are summarized in Table XXXIII. Calculation of one summary point estimate for sensitivity and specificity was not meaningful as both test characteristics shown in Figure 19 were heterogeneous among studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity 0.014). As the Spearman correlation coefficient for sensitivity and specificity was –0.57, it appeared justified to estimate a summary ROC curve as shown in Figure 19.

Accuracy of non-pregnancy prediction

There were also three studies that reported on the capacity of the GAST to predict non-pregnancy after IVF. Sensitivities and specificities for the prediction of non-pregnancy, as calculated from each study, are summarized in Table XXXIV. Again, sensitivity and specificity, as shown in Figure 20, were heterogeneous between studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity 0.005). The Spearman correlation between sensitivity and specificity showed a coefficient of –0.98, sufficient to estimate a summary ROC curve (Figure 20).

Clinical value

Based on the summary ROC curves depicted in Figure 19, a range of positive LRs was calculated and for each ratio, pre-GAST-test probabilities of poor response or non-pregnancy (set at 20 and 80%, respectively) were converted into post-GAST-test probabilities. Table XXXV depicts the probability of obtaining a certain GAST test result and the corresponding LR within different LR ranges for the prediction of poor response and non-pregnancy. At a modest LR of 4–5, the post-GAST-test probability of poor response will not be higher than ~50%, while the chance of obtaining

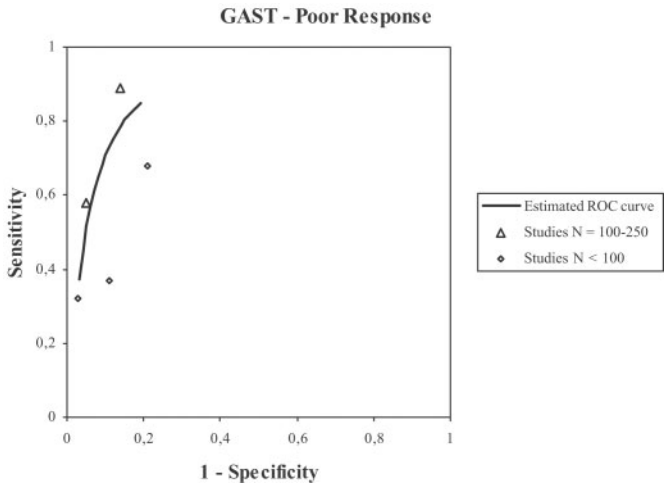

Figure 19. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of the GAST in the prediction of poor response. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

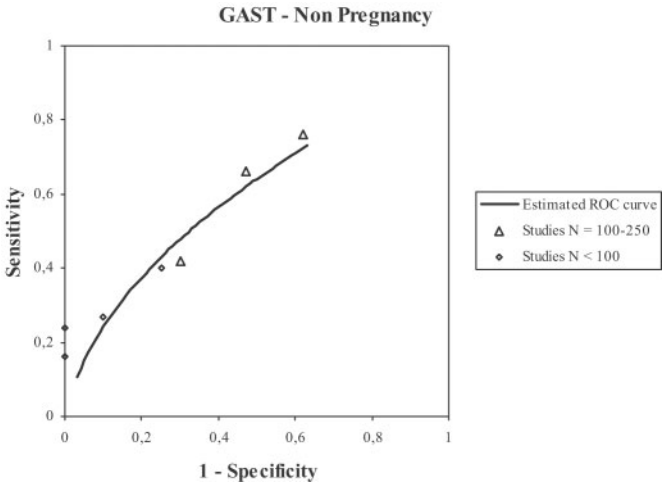

Figure 20. Estimated ROC curve and sensitivity–specificity points for all studies reporting on the performance of the GAST in the prediction of non-pregnancy. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated.

such a test result is quite high, 49%. However, only with an extreme threshold a post-test probability of poor response that approaches 70% can be retained in a considerable number of cases (30%).

For prediction of non-pregnancy, extreme threshold levels are necessary to obtain a modest positive LR of 4–5, leading to a post-test pregnancy rate of approximately 5%. Such abnormal test results occur only in a very limited number of patients, while the false positive rate will lead to unnecessary exclusions from IVF programs if the test is used in a diagnostic fashion.

It can be concluded that with the use of the GAST in regularly cycling women, the accuracy in the prediction of poor response is quite high and could match with those obtained by the use of the AFC. For non-pregnancy prediction the test may only be adequate at a very low threshold level, where hardly any abnormal tests can be found. The results show that the GAST is a candidate for more extensive confirmation research.

Multivariate models

Systematic review

Through the search and selection strategy, a total of nine studies reporting on the predictive capacity of several multivariate models were identified and considered suitable for data extraction and meta-analysis (Balasch *et al*., 1996; Ranieri *et al*., 1998; Creus *et al*., 2000; Fabregues *et al*., 2000; Bancsi *et al*., 2002a; van Rooij *et al*., 2002; Durmusoglu *et al*., 2004; Erdem *et al*., 2004; Muttukrishna *et al*., 2004). Characteristics of the included studies are listed in addendum Table XXXVI. As with most studies on ORTs, definitions for poor response varied considerably. It should be noted that none of the multifactor studies revealed usable data on pregnancy prediction. Moreover, the total number of cases included in these aggregated studies is modest (*n*=991).

Accuracy of poor response prediction

All ten studies only reported on the prediction of poor response. The sensitivities and specificities, the positive LR and the DOR

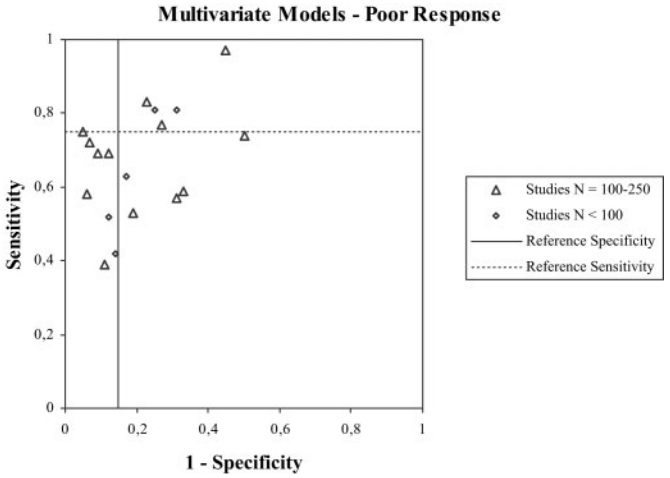

Figure 21. Sensitivity–specificity points for all studies reporting on the performance of multivariate models in the prediction of poor response. Studies reporting on several threshold points are represented by an equivalent number of sens–spec points. Because of heterogeneity among studies no estimated summary ROC point of curve could be constructed. *N* in the legend refers to the number of cycles studied, which in some studies is equivalent to the number of couples treated. Reference lines indicate a desired level for sensitivity (0.75) and specificity (0.85).

for the prediction of poor ovarian response are summarized in Table XXXVII. Calculation of one summary point estimate for sensitivity and specificity did not appear to be possible, as both test characteristics (shown in Figure 21) were heterogeneous among studies (χ^2-test statistic: *P*-value for sensitivity <0.001 and *P*-value for specificity 0.014). As the Spearman correlation coefficient for sensitivity and specificity was –0.45, it appeared unjustified to estimate a summary ROC curve. Regression analysis showed that the performance of one particular test model was not superior to the other, as can also be seen in addendum Table XXXVI from the listing of sensitivities and specificities.

Clinical value

The impossibility of creating summary characteristics makes it difficult to assess the interrelation between positive LR, post-test probability and percentage of abnormal tests. Obviously, clinical value can only be discussed regarding prediction of poor response. It is considered that a challenge test used as a diagnostic tool to identify poor responders should have sensitivity and specificity at a certain desired level. If these levels are set at 75 and 85%, respectively, it can be concluded from Figure 21 that only one study will fulfil these criteria (Bancsi *et al*., 2002a). Especially, the false positive prediction may hamper the use of this test if a high level of detection is needed and patients are refused IVF on the basis of this test. From these data it seems that compared to other ORTs, multifactor models do not create a definite improvement in predictive capacity.

Implications for daily practice

With the postponement of childbearing, the age-related fertility decline has been shown to play an important role in the increase in infertility among couples who are trying to conceive. In IVF treatment, this age effect has been shown in much accumulated data. Because of the variation of female fertility within a certain age category, the need was felt for tests which better identified cases with a state of ovarian reserve that is clearly too low for their age. Because a benchmark for ovarian reserve status in the sense of quantity and quality is lacking, the occurrence of poor ovarian response to maximal stimulation and the occurrence of pregnancy in IVF are used as parameters to assess the accuracy of the test. The ideal ORT should identify a substantial percentage of IVF-indicated cases which have a practically zero chance of becoming pregnant because of the adverse effects of diminished ovarian reserve in a series of treatment cycles. Those cases can be refrained from entering the programme, as the very high costs involved will have only minimal results. If not too expensive and not too demanding for the patient, such a test would be readily embraced by physicians, patients, health politicians and insurance companies.

From the systematic and meta-analytic reviews presented here, it can be concluded that the ORTs known to date have only very modest predictive properties and are therefore far from suitable for relevant clinical use. Although mostly cheap and not very demanding, their accuracy, especially in the prediction of the occurrence of pregnancy, is very limited. If a high threshold is used, to prevent couples from wrongly being refused IVF, a very small minority of IVF-indicated cases (~3%) were identified as having unfavourable prospects in an IVF treatment cycle (pregnancy rate for that cycle = 5%). It should be noted that the use of pregnancy as outcome parameter for the assessment of ovarian reserve status may be insufficient if only one exposure cycle is taken into account. As such, the possibility of misjudgement on the basis of currently known ORTs is hard to rule out. This implies that the use of the test as a method to deny treatment to assumed ovarian aged women should be declined and, as a consequence the test should not be applied on a regular basis and should only be used for counselling or screening purposes.

Accuracy of testing for the occurrence of poor ovarian response to stimulation appeared to be clearly better than for the occurrence of pregnancy. This may be understood in the light of the following factors: (i) that the chance of pregnancy after IVF depends on many more factors than ovarian reserve alone, (ii) that the occurrence of pregnancy after an ORT was usually evaluated in only one IVF cycle and as such may not accurately represent a female's true reproductive capacity and (iii) that the response to stimulation is likely to represent the size of the cohort of FSH-sensitive follicles continuously present in the ovaries and is directly related to the magnitude of ovarian reserve (i.e. the remaining primordial follicle pool (Gougeon, 1984). Poor ovarian response has been associated with a reduced chance of pregnancy in the actual treatment cycle as well as in subsequent cycles and as such may well be indicative of ovarian reserve status in both the quantitative and qualitative sense (Ulug *et al*., 2003; Klinkert *et al*., 2004; Klinkert *et al*., 2005a). Accurate prediction of poor response could therefore have clinical value if the pregnancy prospects are so unfavourable that a predicted poor responder would be denied treatment. Accuracy in response prediction, however, will only be high if the false positives are prevented by using extreme threshold levels, implicating that only minor percentages of abnormal tests will be found and many future poor responders will pass unrecognized. At the same time it is necessary to know whether the predicted poor responder indeed has very low prospects for

success in subsequent cycles. As much of this is unknown at the present time, the use of any ORT for poor response prediction cannot be supported, not even if it would be used for adapting the treatment schedule in anticipated poor responders, as an altered treatment schedule has consistently been shown to be effective in women with a severely reduced size of follicle cohort (Tarlatzis *et al*., 2003; Klinkert, 2005; Klinkert *et al*., 2005a).

One aspect of clinical value deserves some special attention. ORTs are mostly used as a diagnostic test, indicating that in case of an abnormal test result, the diagnosis that there is diminished ovarian reserve is made (Scott and Hofmann, 1995; Levi *et al*., 2001). From the fact that for evaluation of the test, proxy variables of true ovarian reserve (poor ovarian response and non-pregnancy) are used and that false positive test results may eliminate couples from the IVF trail even if they do have adequate prospects, it becomes clear that ORTs may better be considered as *screening tests*. All this implies that an abnormal test necessitates confirmation by another test. This other test may, for instance, be a first IVF attempt where ovarian response is the additional test. Alternatively, combinations of independent predictive tests or repeating of the initial test could improve the diagnostic performance of the single test (Ng *et al*., 2000; Bancsi *et al*., 2002b; van Rooij *et al*., 2002; Popovic-Todorovic *et al*., 2003a,b; Bancsi *et al*., 2004a,b).

As poor ovarian response will provide some information on ovarian reserve status, especially if the stimulation is maximal, entering the first cycle of IVF without any prior testing seems to be the preferable strategy. Once a poor response is obtained, the question arises whether this finding is based on depleted ovaries or other causes, like underdosing for instance, based on the presence of certain FSH receptor polymorphisms (Perez *et al*., 2000; Behre *et al*., 2005; Greb *et al*., 2005; de Koning *et al*., 2005). A repeat cycle with adequate, maximal stimulation or a *post hoc*-performed ORT [basal FSH or AFC (Hendriks *et al*., 2005c)] may correctly classify the poor responder patient as having an aged ovary and may correctly suggest that they refrain from further treatment (Klinkert *et al*., 2004).

It should be remembered that the purpose of any ORT is the identification of women with poor ovarian reserve for their age. This implies that chronological age always is the first step in ovarian reserve assessment. In young women, ORTs may help to classify poor responders and in direct management in these cases by estimating the size of the FSH-sensitive cohort. In older women, ORTs may help to identify those cases that, in spite of their age, still may have acceptable chances of becoming pregnant through IVF as the quantity of response to stimulation is anticipated to be normal or even high (Klinkert *et al*., 2005b).

Future perspectives in this research field may be found in studies where success rates in cumulative treatment cycles or in units of time (1-year treatment periods) are analysed to answer the question of whether any test will correctly identify those couples who will not become pregnant in such series of exposures. Novel tests that most accurately estimate the age at which menopause is expected to take place in an individual woman may facilitate the estimation of the remaining reproductive potential at a certain age. Such tests will probably be based on family history (age at menopause of mother) or will comprise testing for genetic markers, which may be discovered from large-scale population genetic screening.

References

Abma JC, Chandra A, Mosher WD, Peterson LS and Piccinino LJ (1997) Fertility, family planning, and women's health: new data from the 1995 National Survey of Family Growth. Vital Health Stat 23,1–114.

Akande VA, Keay SD, Hunt LP, Mathur RS, Jenkins JM and Cahill DJ (2004) The practical implications of a raised serum FSH and age on the risk of IVF treatment cancellation because of a poor ovarian response. J Assist Reprod Genet 21,257–262.

Balasch J, Creus M, Fabregues F, Carmona F, Casamitjana R, Ascaso C and Vanrell JA (1996) Inhibin, follicle-stimulating hormone, and age as predictors of ovarian response in in vitro fertilization cycles stimulated with gonadotropin-releasing hormone agonist-gonadotropin treatment. Am J Obstet Gynecol 175,1226–1230.

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.

Bancsi LF, Broekmans FJ, Eijkemans MJ, de Jong FH, Habbema JD, te Velde ER (2002a) Predictors of poor ovarian response in in vitro fertilization: a prospective study comparing basal markers of ovarian reserve. Fertil Steril 77,328–336.

Bancsi LFJMM, Broekmans FJM, Eijkemans MJC, de Jong FH, Habbema JDF and te Velde ER (2002b) Predictors of poor ovarian response in in vitro fertilization: a prospective study comparing basal markers of ovarian reserve. Fertil Steril 77,328–336.

Bancsi LF, Broekmans FJ, Mol BW, Habbema JD and te Velde ER (2003) Performance of basal follicle-stimulating hormone in the prediction of poor ovarian response and failure to become pregnant after in vitro fertilization: a meta-analysis. Fertil Steril 79,1091–1100.

Bancsi LF, Broekmans FJ, Looman CW, Habbema JD and te Velde ER (2004a) Impact of repeated antral follicle counts on the prediction of poor ovarian response in women undergoing in vitro fertilization. Fertil Steril 81,35–41.

Bancsi LF, Broekmans FJ, Looman CW, Habbema JD and te Velde ER (2004b) Predicting poor ovarian response in IVF: use of repeat basal FSH measurement. J Reprod Med 49,187–194.

Bassil S, Godin PA, Gillerot S, Verougstraete JC and Donnez J (1999) In vitro fertilization outcome according to age and follicle-stimulating hormone levels on cycle day 3. J Assist Reprod Genet 16,236–241.

Behre HM, Greb RR, Mempel A, Sonntag B, Kiesel L, Kaltwasser P, Seliger E, Ropke F, Gromoll J, Nieschlag E and Simoni M (2005) Significance of a common single nucleotide polymorphism in exon 10 of the follicle-stimulating hormone (FSH) receptor gene for the ovarian response to FSH: a pharmacogenetic approach to controlled ovarian hyperstimulation. Pharmacogenet Genomics 15,451–456.

Block E (1952) Quantitative morphological investigations of the follicular system in women. Variations at different ages. Acta Anat (Basel) 14 (Suppl 16),108–123.

de Boer EJ, den TI, te Velde ER, Burger CW, Klip H and van Leeuwen FE (2002) A low number of retrieved oocytes at in vitro fertilization treatment is predictive of early menopause. Fertil Steril 77,978–985.

de Bruin JP and te Velde ER (2004) Female reproductive ageing: concepts and consequences. In Tulandi T and Gosden RG (eds) *Preservation of Fertility*. London, UK: Taylor & Francis, p. 3.

Chae HD, Kim CH, Kang BM and Chang YS (2000) Clinical usefulness of basal FSH as a prognostic factor in patients undergoing intracytoplasmic sperm injection. J Obstet Gynaecol Res 26,55–60.

Chan YF, Ho PC, So WW and Yeung WS (1993) Basal serum pituitary hormone levels and outcome of in vitro fertilization utilizing a flare nasal gonadotropin releasing hormone agonist and fixed low-dose follicle-stimulating hormone/human menopausal gonadotropin regimen. J Assist Reprod Genet 10,251–254.

Chang MY, Chiang CH, Chiu TH, Hsieh TT, Soong YK (1998a). The antral follicle count predicts the outcome of pregnancy in a controlled ovarian hyperstimulation/intrauterine insemination program. J Assist Reprod Genet 15,12–17.

Chang MY, Chiang CH, Hsieh TT, Soong YK and Hsu KH (1998b) Use of the antral follicle count to predict the outcome of assisted reproductive technologies. Fertil Steril 69,505–510.

Chuang CC, Chen CD, Chao KH, Chen SU, Ho HN and Yang YS (2003) Age is a better predictor of pregnancy potential than basal follicle-stimulating hormone levels in women undergoing in vitro fertilization. Fertil Steril 79,63–68.

Collins JA, Burrows EA and Wilan AR (1995) The prognosis for live birth among untreated infertile couples. Fertil Steril 64,22–28.

Creus M, Penarrubia J, Fabregues F, Vidal E, Carmona F, Casamitjana R, Vanrell JA and Balasch J (2000) Day 3 serum inhibin B and FSH and age as predictors of assisted reproduction treatment outcome. Hum Reprod 15,2341–2346.

Csemiczky G, Wramsby H and Landgren BM (1996) Luteal phase oestradiol and progesterone levels are stronger predictors than follicular phase follicle stimulating hormone for the outcome of in-vitro fertilization treatment in women with tubal infertility. Hum Reprod 11,2396–2399.

Csemiczky G, Harlin J and Fried G (2002) Predictive power of clomiphene citrate challenge test for failure of in vitro fertilization treatment. Acta Obstet Gynecol Scand 81,954–961.

Deeks JJ (2001) Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. BMJ 323,157–162.

Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA and Bezemer PD (2002) Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2,9.

Durmusoglu F, Elter K, Yoruk P and Erenus M (2004) Combining cycle day 7 follicle count with the basal antral follicle count improves the prediction of ovarian response. Fertil Steril 81,1073–1078.

Ebrahim A, Rienhardt G, Morris S, Kruger TF, Lombard CJ and Van der Merwe JP (1993) Follicle stimulating hormone levels on cycle day 3 predict ovulation stimulation response. J Assist Reprod Genet 10,130–136.

Eimers JM, te Velde ER, Gerritse R, Vogelzang ET, Looman CW and Habbema JD (1994). The prediction of the chance to conceive in subfertile couples. Fertil Steril 61,44–52.

Engmann L, Sladkevicius P, Agrawal R, Bekir J, Campbell S and Tan SL (1999a) The pattern of changes in ovarian stromal and uterine artery blood flow velocities during in vitro fertilization treatment and its relationship with outcome of the cycle. Ultrasound Obstet Gynecol 13,26–33.

Engmann L, Sladkevicius P, Agrawal R, Bekir JS, Campbell S and Tan SL (1999b) Value of ovarian stromal blood flow velocity measurement after pituitary suppression in the prediction of ovarian responsiveness and outcome of in vitro fertilization treatment. Fertil Steril 71,22–29.

Erdem A, Erdem M, Biberoglu K, Hayit O, Arslan M and Gursoy R (2002) Age-related changes in ovarian volume, antral follicle counts and basal FSH in women with normal reproductive health. J Reprod Med 47,835–839.

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.

Esposito MA, Coutifaris C and Barnhart KT (2002) A moderately elevated day 3 FSH concentration has limited predictive value, especially in younger women. Hum Reprod 17,118–123.

Evers JL, Slaats P, Land JA, Dumoulin JC and Dunselman GA (1998) Elevated levels of basal estradiol-17beta predict poor response in patients with normal basal levels of follicle-stimulating hormone undergoing in vitro fertilization. Fertil Steril 69,1010–1014.

Fabregues F, Balasch J, Creus M, Carmona F, Puerto B, Quinto L, Casamitjana R and Vanrell JA (2000) Ovarian reserve test with human menopausal gonadotropin as a predictor of in vitro fertilization outcome. J Assist Reprod Genet 17,13–19.

Fanchin R, de Ziegler D, Olivennes F, Taieb J, Dzik A and Frydman R (1994) Exogenous follicle stimulating hormone ovarian reserve test (EFORT): a simple and reliable screening test for detecting 'poor responders' in in-vitro fertilization. Hum Reprod 9,1607–1611.

Fasouliotis SJ, Simon A and Laufer N (2000) Evaluation and treatment of low responders in assisted reproductive technology: a challenge to meet. J Assist Reprod Genet 17,357–373.

Fiçicioğlu C, Kutlu T, Demirbasoglu S and Mulayim B (2003) The role of inhibin B as a basal determinant of ovarian reserve. Gynecol Endocrinol 17,287–293.

Fisch JD and Sher G (2002) The antral follicle count (AFC) correlates with the metaphase II oocytes and ART cycle outcome: an update. Fertil Steril 78,S90.

Frattarelli JL, Lauria-Costab DF, Miller BT, Bergh PA and Scott RT (2000) Basal antral follicle number and mean ovarian diameter predict cycle cancellation and ovarian responsiveness in assisted reproductive technology cycles. Fertil Steril 74,512–517.

Frattarelli JL, Levi AJ, Miller BT and Segars JH (2003) A prospective assessment of the predictive value of basal antral follicles in in vitro fertilization cycles. Fertil Steril 80,350–355.

Glas AS, Lijmer JG, Prins MH, Bonsel GJ and Bossuyt PM (2003) The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 56,1129–1135.

Gougeon A (1984) Caracteres qualitatifs et quantitatifs de la population folliculaire dans l'ovaire humaine adulte. Contracept Fertil Sex 12(3),527–535.

Greb RR, Grieshaber K, Gromoll J, Sonntag B, Nieschlag E, Kiesel L and Simoni M (2005) A common single nucleotide polymorphism in exon 10 of the human follicle stimulating hormone receptor is a major determinant of length and hormonal dynamics of the menstrual cycle. J Clin Endocrinol Metab 90,4866–4872.

Grimes DA and Schulz KF (2005) Refining clinical diagnosis with likelihood ratios. Lancet 365,1500–1505.

Gulekli B, Bulbul Y, Onvural A, Yorukoglu K, Posaci C, Demir N and Erten O (1999) Accuracy of ovarian reserve tests. Hum Reprod 14,2822–2826.

Gurgan T, Urman B, Yarali H and Duran HE (1997) Follicle-stimulating hormone levels on cycle day 3 to predict ovarian response in women undergoing controlled ovarian hyperstimulation for in vitro fertilization using a flare-up protocol. Fertil Steril 68,483–487.

Hall JE, Welt CK and Cramer DW (1999) Inhibin A and inhibin B reflect ovarian function in assisted reproduction but are less useful at predicting outcome. Hum Reprod 14,409–415.

Hazout A, Bouchard P, Seifer DB, Aussage P, Junca AM and Cohen-Bacrie P (2004) Serum antimullerian hormone/mullerian-inhibiting substance appears to be a more discriminatory marker of assisted reproductive technology outcome than follicle-stimulating hormone, inhibin B, or estradiol. Fertil Steril 82,1323–1329.

Hendriks DJ, Broekmans FJ, Bancsi LF, de Jong FH, Looman CW and te Velde ER (2005a) Repeated clomiphene citrate challenge testing in the prediction of outcome in IVF: a comparison with basal markers for ovarian reserve. Hum Reprod 20,163–169.

Hendriks DJ, Broekmans FJ, Bancsi LF, Looman CW, de Jong FH and te Velde ER (2005b) Single and repeated GnRH agonist stimulation tests compared with basal markers of ovarian reserve in the prediction of outcome in IVF. J Assist Reprod Genet 22,65–73.

Hendriks DJ, Mol BW, Bancsi LF, te Velde ER and Broekmans FJ (2005c) Antral follicle count in the prediction of poor ovarian response and pregnancy after in vitro fertilization: a meta-analysis and comparison with basal follicle-stimulating hormone level. Fertil Steril 83,291–301.

Hendriks DJ, te Velde ER, Looman CW, Bancsi LF and Broekmans FJ (2005d). The role of poor response in the prediction of the cumulative ongoing pregnancy rate in in vitro fertilisation. Dynamic and basal ovarian reserve tests for outcome prediction in IVF: comparisons and meta-analyses. Academic Thesis, Utrecht, 162–179.

Honest H and Khan KS (2002) Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC Health Serv Res 2,4.

Hsieh YY, Chang CC and Tsai HD (2001) Antral follicle counting in predicting the retrieved oocyte number after ovarian hyperstimulation. J Assist Reprod Genet 18,320–324.

Hunault CC, Habbema JD, Eijkemans MJ, Collins JA, Evers JL and te Velde ER (2004) Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models. Hum Reprod 19,2019–2026.

Hunault CC, Laven JS, van Rooij IA, Eijkemans MJ te Velde ER and Habbema JD (2005) Prospective validation of two models predicting pregnancy leading to live birth among untreated subfertile couples. Hum Reprod 20,1636–1641.

Huyser C, Fourie FL, Pentz J and Hurter P (1995) The predictive value of basal follicle stimulating and growth hormone levels as determined by immunofluorometry during assisted reproduction. J Assist Reprod Genet 12,244–251.

Irwig L, Macaskill P, Glasziou P and Fahey M (1995) Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 48,119–130.

Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC and Mosteller F (1994) Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 120,667–676.

Jain T, Soules MR and Collins JA (2004) Comparison of basal follicle-stimulating hormone versus the clomiphene citrate challenge test for ovarian reserve screening. Fertil Steril 82,180–185.

Jarvela IY, Sladkevicius P, Kelly S, Ojha K, Campbell S and Nargund G (2003) Quantification of ovarian power Doppler signal with three-dimensional ultrasonography to predict response during in vitro fertilization. Obstet Gynecol 102,816–822.

Jinno M, Hoshiai T, Nakamura Y, Teruya K and Tsunoda T (2000) A novel method for assessing assisted female fertility: bioelectric impedance. J Clin Endocrinol Metab 85,471–474.

Kahraman S, Vicdan K, Isik AZ, Ozgun OD, Alaybeyoglu L, Polat G and Biberoglu K (1997) Clomiphene citrate challenge test in the assessment of ovarian reserve before controlled ovarian hyperstimulation for intracytoplasmic sperm injection. Eur J Obstet Gynecol Reprod Biol 73, 177–182.

Khalifa E, Toner JP, Muasher SJ, Acosta AA (1992). Significance of basal follicle-stimulating hormone levels in women with one ovary in a program of in vitro fertilization. Fertil Steril 57,835–839.

Kim SH, Ku SY, Jee BC, Suh CS, Moon SY and Lee JY (2002) Clinical significance of transvaginal color Doppler ultrasonography of the ovarian artery as a predictor of ovarian response in controlled ovarian hyperstimulation for in vitro fertilization and embryo transfer. J Assist Reprod Genet 19,103–112.

Klinkert ER (2005) Clinical significance and management of poor response in IVF. Academic Thesis, Utrecht.

Klinkert ER, Broekmans FJ Looman CW and te Velde ER (2004) A poor response in the first in vitro fertilization cycle is not necessarily related to a poor prognosis in subsequent cycles. Fertil Steril 81,1247–1253.

Klinkert ER, Broekmans FJ, Looman CW, Habbema JD and te Velde ER (2005a) Expected poor responders on the basis of an antral follicle count do not benefit from a higher starting dose of gonadotrophins in IVF treatment: a randomized controlled trial. Hum Reprod 20,611–615.

Klinkert ER, Broekmans FJ, Looman CW, Habbema JD and te Velde ER (2005b) 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.

de Koning CH, Benjamins T, Harms P, Homburg R, van Montfrans JM, Gromoll J, Simoni M and Lambalk CB (2006) The distribution of FSH receptor isoforms is related to basal FSH levels in subfertile women with normal menstrual cycles. Hum Reprod 21,443–446.

van Kooij RJ, Looman CW, Habbema JD, Dorland M and te Velde ER (1996) Age-dependent decrease in embryo implantation rate after in vitro fertilization. Fertil Steril 66,769–775.

Kupesic S and Kurjak A (2002) Predictors of IVF outcome by three-dimensional ultrasound. Hum Reprod 17,950–955.

Kupesic S, Kurjak A, Bjelos D and Vujisic S (2003) Three-dimensional ultrasonographic ovarian measurements and in vitro fertilization outcome are related to age. Fertil Steril 79,190–197.

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.

Lambalk CB, de Koning CH, Flett A, van Kasteren Y, Gosden R and Homburg R (2004) Assessment of ovarian reserve. Ovarian biopsy is not a valid method for the prediction of ovarian reserve. Hum Reprod 19,1055–1059.

Lass A (2001) Assessment of ovarian reserve – is there a role for ovarian biopsy? Hum Reprod 16,1055–1057.

Lass A (2004) Assessment of ovarian reserve: is there still a role for ovarian biopsy in the light of new data? Hum Reprod 19,467–469.

Lass A, Silye R, Abrams DC, Krausz T, Hovatta O, Margara R and Winston RM (1997a) Follicular density in ovarian biopsy of infertile women: a novel method to assess ovarian reserve. Hum Reprod 12,1028–1031.

Lass A, Skull J, McVeigh E, Margara R and Winston RM (1997b) Measurement of ovarian volume by transvaginal sonography before ovulation induction with human menopausal gonadotrophin for in- vitro fertilization can predict poor response. Hum Reprod 12,294–297.

Lawson R, El Toukhy T, Kassab A, Taylor A, Braude P, Parsons J and Seed P (2003) Poor response to ovulation induction is a stronger predictor of early menopause than elevated basal FSH: a life table analysis. Hum Reprod 18,527–533.

Leridon H (1998) [30 years of contraception in France]. Contracept Fertil Sex 26,435–438.

Levi AJ, Raynault MF, Bergh PA, Drews MR, Miller BT and Scott RT (2001) Reproductive outcome in patients with diminished ovarian reserve. Fertil Steril 76,666–669.

Licciardi FL, Liu HC and Rosenwaks Z (1995) Day 3 estradiol serum concentrations as prognosticators of ovarian stimulation response and pregnancy outcome in patients undergoing in vitro fertilization. Fertil Steril 64,991–994.

Littenberg B and Moses LE (1993) Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 13,313–321.

Loumaye E, Billion JM, Mine JM, Psalti I, Pensis M and Thomas K (1990) Prediction of individual response to controlled ovarian hyperstimulation by means of a clomiphene citrate challenge test. Fertil Steril 53,295–301.

Martin JS, Nisker JA, Tummon IS, Daniel SA, Auckland JL and Feyles V (1996) Future in vitro fertilization pregnancy potential of women with variably elevated day 3 follicle-stimulating hormone levels. Fertil Steril 65,1238–1240.

Midgette AS, Stukel TA and Littenberg B (1993) A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. Med Decis Making 13,253–257.

Mikkelsen AL, Andersson AM, Skakkebaek NE and Lindenberg S (2001) Basal concentrations of oestradiol may predict the outcome of in-vitro maturation in regularly menstruating women. Hum Reprod 16,862–867.

Mol BW, Dijkman B, Wertheim P, Lijmer J, van d V and Bossuyt PM (1997) The accuracy of serum chlamydial antibodies in the diagnosis of tubal pathology: a meta-analysis. Fertil Steril 67,1031–1037.

Moses LE, Shapiro D and Littenberg B (1993) Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 12,1293–1316.

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.

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.

Nahum R, Shifren JL, Chang Y, Leykin L, Isaacson K and Toth TL (2001) Antral follicle assessment as a tool for predicting outcome in IVF – is it a better predictor than age and FSH? J Assist Reprod Genet 18,151–155.

National Collaborating Center for Women's and Children's Health. (2004) *Fertility: Assessment and Treatment for People with Fertility Problems*. RCOG press, UK.

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.

Padilla SL, Bayati J and Garcia JE (1990) Prognostic value of the early serum estradiol response to leuprolide acetate in in vitro fertilization. Fertil Steril 53,288–294.

Penarrubia J, Balasch J, Fabregues F, Carmona F, Casamitjana R, Moreno V, Calafell JM and Vanrell JA (2000) Day 5 inhibin B serum concentrations as predictors of assisted reproductive technology outcome in cycles stimulated with gonadotrophin-releasing hormone agonist-gonadotrophin treatment. Hum Reprod 15,1499–1504.

Penarrubia J, Fabregues F, Manau D, Creus M, Casals G, Casamitjana R, Carmona F, Vanrell JA and Balasch J (2005) Basal and stimulation day 5 anti-Mullerian hormone serum concentrations as predictors of ovarian response and pregnancy in assisted reproductive technology cycles stimulated with gonadotropin-releasing hormone agonist – gonadotropin treatment. Hum Reprod 20,915–922.

Perez MM, Gromoll J, Behre HM, Gassner C, Nieschlag E and Simoni M (2000) Ovarian response to follicle-stimulating hormone (FSH) stimulation depends on the FSH receptor genotype. J Clin Endocrinol Metab 85,3365–3369.

Phophong P, Ranieri DM, Khadum I, Meo F and Serhal P (2000) Basal 17beta-estradiol did not correlate with ovarian response and in vitro fertilization treatment outcome. Fertil Steril 74,1133–1136.

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.

Popovic-Todorovic B, Loft A, Lindhard A, Bangsboll S, Andersson AM and Andersen AN (2003b) 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.

Anonymous (1995) Pregnancies and births resulting from in vitro fertilization: French national registry analysis of data 1986 to 1990. FIVNAT (French In Vitro National). Fertil Steril 64,746–756.

Pruksananonda K, Boonkasemsanti W and Virutamasen P (1996) Basal follicle – stimulating hormone levels on day 3 of previous cycle are predictive of in vitro fertilization outcome. J Med Assoc Thai 79,365–369.

Qu J, Godin PA, Nisolle M and Donnez J (2000) Distribution and epidermal growth factor receptor expression of primordial follicles in human ovarian tissue before and after cryopreservation. Hum Reprod 15,302–310.

Ranieri DM, Quinn F, Makhlouf A, Khadum I, Ghutmi W, McGarrigle H, Davies M and Serhal P (1998) Simultaneous evaluation of basal follicle-stimulating hormone and 17 beta-estradiol response to gonadotropin-releasing hormone analogue stimulation: an improved predictor of ovarian reserve. Fertil Steril 70,227–233.

Ranieri DM, Phophong P, Khadum I, Meo F, Davis C and Serhal P (2001) Simultaneous evaluation of basal FSH and oestradiol response to GnRH analogue (F-G-test) allows effective drug regimen selection for IVF. Hum Reprod 16,673–675.

Roberts JE, Spandorfer S, Fasouliotis SJ, Kashyap S and Rosenwaks Z (2005) Taking a basal follicle-stimulating hormone history is essential before initiating in vitro fertilization. Fertil Steril 83,37–41.

van Rooij IAJ, Broekmans FJ, te Velde ER, Fauser BCJM, Bancsi LFJMM, de Jong FH and Themmen APN (2002) Serum anti-Mullerian hormone levels: a novel measure of ovarian reserve. Hum Reprod 17,101–107.

van Rooij IA, Broekmans FJ, Hunault CC, Scheffer GJ, Eijkemans MJ, de Jong FH, Themmen AP and te Velde ER (2006) The use of ovarian reserve tests for the prediction of ongoing pregnancy in couples with unexplained female subfertility. Reprod Biomed Online 12, 182–190.

Schild R L, Knobloch C, Dorn C, Fimmers R, van d V and Hansmann M (2001). The role of ovarian volume in an in vitro fertilization programme as assessed by 3D ultrasound. Arch Gynecol Obstet 265,67–72.

Schmidt KL, Ernst E, Byskov AG, Nyboe AA and Yding AC (2003) Survival of primordial follicles following prolonged transportation of ovarian tissue prior to cryopreservation. Hum Reprod 18,2654–2659.

Scott RT Jr and Hofmann GE (1995) Prognostic assessment of ovarian reserve [see comments]. Fertil Steril 63,1–11.

Scott RT, Toner JP, Muasher SJ, Oehninger S, Robinson S and Rosenwaks Z (1989) Follicle-stimulating hormone levels on cycle day 3 are predictive of in vitro fertilization outcome. Fertil Steril 51,651–654.

Seifer DB, Lambert Messerlian G, Hogan JW, Gardiner AC, Blazar AS and Berk CA (1997) Day 3 serum inhibin-B is predictive of assisted reproductive technologies outcome [see comments]. Fertil Steril 67,110–114.

Sharara FI and McClamrock HD (1999) The effect of aging on ovarian volume measurements in infertile women. Obstet Gynecol 94,57–60.

Sharara FI and McClamrock HD (2000) Antral follicle count and ovarian volume predict IVF outcome. Fertil Steril 74,S176.

Sharara FI and Scott RT (2004) Assessment of ovarian reserve. Is there still a role for ovarian biopsy? First do no harm! Hum Reprod 19,470–471.

Sharif K, Elgendy M, Lashen H and Afnan M (1998) Age and basal follicle stimulating hormone as predictors of in vitro fertilisation outcome. Br J Obstet Gynaecol 105,107–112.

Sharma V, Allgar V and Rajkhowa M (2002) Factors influencing the cumulative conception rate and discontinuation of in vitro fertilization treatment for infertility. Fertil Steril 78,40–46.

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.

Smotrich DB, Widra EA, Gindoff PR, Levy MJ, Hall JL and Stillman RJ (1995) Prognostic value of day 3 estradiol on in vitro fertilization outcome. Fertil Steril 64,1136–1140.

Snick HK, Snick TS, Evers JL and Collins JA (1997) The spontaneous pregnancy prognosis in untreated subfertile couples: the Walcheren primary care study. Hum Reprod 12,1582–1588.

Spira A (1988) The decline of fecundity with age. Raturitas 10 (Suppl),15–22.

van der Stege JG and van der Linden PJ (2001) Useful predictors of ovarian stimulation response in women undergoing in vitro fertilization. Gynecol Obstet Invest 52,43–46.

Stolwijk AM, Zielhuis GA, Hamilton CJ, Straatman H, Hollanders JM, Goverde HJ, van Dop PA and Verbeek AL (1996) Prognostic models for the probability of achieving an ongoing pregnancy after in-vitro fertilization and the importance of testing their predictive value. Hum Reprod 11,2298–2303.

Stolwijk AM, Straatman H, Zielhuis GA, Jansen CA, Braat DD, van Dop PA and Verbeek AL (1998) External validation of prognostic models for ongoing pregnancy after in-vitro fertilization. Hum Reprod 13,3542–3549.

Syrop CH, Willhoite A and Van Voorhis BJ (1995) Ovarian volume: a novel outcome predictor for assisted reproduction. Fertil Steril 64,1167–1171.

Tanbo T, Dale PO, Abyholm T and Stokke KT (1989) Follicle-stimulating hormone as a prognostic indicator in clomiphene citrate/human menopausal gonadotrophin-stimulated cycles for in-vitro fertilization. Hum Reprod 4,647–650.

Tanbo T, Abyholm T, Bjoro T and Dale PO (1990) Ovarian stimulation in previous failures from in-vitro fertilization: distinction of two Groups of poor responders. Hum Reprod 5,811–815.

Tanbo T, Dale PO, Lunde O, Norman N and Abyholm T (1992) Prediction of response to controlled ovarian hyperstimulation: a comparison of basal and clomiphene citrate-stimulated follicle- stimulating hormone levels. Fertil Steril 57,819–824.

Tarlatzis BC, Zepiridis L, Grimbizis G and Bontis J (2003) Clinical management of low ovarian response to stimulation for IVF: a systematic review. Hum Reprod Update 9,61–76.

Templeton A, Morris JK and Parslow W (1996) Factors that affect outcome of in-vitro fertilisation treatment [see comments]. Lancet 348,1402–1406.

Toner JP, Philput CB, Jones GS and Muasher SJ (1991) Basal follicle-stimulating hormone level is a better predictor of in vitro fertilization performance than age. Fertil Steril 55,784–791.

Ulug U, Ben Shlomo I, Turan E, Erden HF, Akman MA and Bahceci M (2003) Conception rates following assisted reproduction in poor responder patients: a retrospective study in 300 consecutive cycles. Reprod Biomed Online 6,439–443.

Vazquez ME, Verez JR, Stern JJ, Gutierrez NA and Asch RH (1998) Elevated basal estradiol levels have no negative prognosis in young women undergoing ART cycles. Gynecol Endocrinol 12,155–159.

te Velde ER and Pearson PL (2002) The variability of female reproductive aging. Hum Reprod Update 8,141–154.

Ventura SJ, Mosher WD, Curtin SC, Abma JC and Henshaw S (2001) Trends in pregnancy rates for the United States 1976–97: an update. Natl Vital Stat Rep 49,1–9.

Webber LJ, Stubbs S, Stark J, Trew GH, Margara R, Hardy K and Franks S (2003) Formation and early development of follicles in the polycystic ovary. Lancet 362,1017–1021.

Weinstein M, Wood AJ and Chang MC (1993) Age patterns in fecundability. In Gray R, Leridon H and Spira A (eds) *Biomedical and Demographic Determinants of Reproduction*. Clarendon Press, Oxford, pp. 209–220.

Winslow KL, Toner JP, Brzyski RG, Oehninger SC, Acosta AA and Muasher SJ (1991) The gonadotropin-releasing hormone agonist stimulation test – a sensitive predictor of performance in the flare-up in vitro fertilization cycle. Fertil Steril 56,711–717.

Wood JW (1989) Fecundity and natural fertility in humans. Oxf Rev Reprod Biol 11,61–109.

Yanushpolsky EH, Hurwitz S, Tikh E and Racowsky C (2003) Predictive usefulness of cycle day 10 follicle-stimulating hormone level in a clomiphene citrate challenge test for in vitro fertilization outcome in women younger than 40 years of age. Fertil Steril 80,111–115.

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.

Zaidi J, Barber J, Kyei Mensah A, Bekir J, Campbell S and Tan SL (1996). Relationship of ovarian stromal blood flow at the baseline ultrasound scan to subsequent follicular response in an in vitro fertilization program. Obstet Gynecol 88,779–784.

Submitted on December 4, 2005; resubmitted on April 27, 2006; accepted on June 12, 2006

Addendum

Table II. Characteristics of included studies on Basal FSH (computerized search using the test-specific keywords *follicle stimulating hormone and FSH*)

Author	Consecutive	One cycle per couple	Data per	Definition		FSH-assay
				Poor response/Cancel	Pregnancy	
Scott *et al.*	Yes	No	Cycle	Not stated	Clinical/ongoing	RIA: Leeco Diagnostics
Padilla *et al.*	No	No	Cycle	Not stated	Clinical	RIA: Amersham Corp.
Toner *et al.*	No	No	Cycle/retrieval	<2 follicles 16 mm	Ongoing	RIA: Leeco Diagnostics
Khalifa *et al.*	No	No	Retrieval	Not stated	Ongoing	RIA: Leeco Diagnostics
Ebrahim *et al.*	Yes	Yes	Cycle	<3 oocytes	Term	RIA: Serono Diagnostics
Chan *et al.*	No	Not stated	Cycle	<3 follicles 15 mm	Clinical/ongoing	RIA: Diag. Products Inc.
Fanchin *et al.*	Yes	Yes	Cycle	<3 oocytes	Not applicable	Immunometric: Kodak Diag.
Huyser *et al.*	No	Yes	Cycle	Not stated	Term	IFMA: Delfia
Licciardi *et al.*	No	Not stated	Retrieval	Not stated	Ongoing	RIA: Leeco Diagnostics
Smotrich *et al.*	No	No	Cycle	<2 follicles 16 mm	Clinical	RIA: Nichols Inst. Radio.
Martin *et al.*	Yes	No	Cycle	Not stated	Clinical	ACS-180: Chemilum.
Pruksanonda *et al.*	No	Yes	Cycle	<3 follicle	Clinical	Fuorescense immunoassay
Csemiczky *et al.*	No	No	Cycle	Not stated	Clinical	RIA: Diag. Products Inc.
Balasch *et al.*	Yes	Yes	Cycle	<2 follicles 17 mm. or <5 follicles 14 mm	Not applicable	RIA: Immunotech Int.
Gurgan *et al.*	No	Yes	Cycle	<2 follicles 18 mm	Clinical	RIA: J&J Clin. Diagnostics
Sharif *et al.*	Yes	Yes	Cycle	<4 follicles 14 mm	Clinical	ACS-180: Chemilum.
Chang *et al.*	Yes	No	Cycle	Not stated	Clinical	Not stated
Evers *et al.*	Yes	Yes	Cycle	<4 follicles 14 mm	Clinical	RIA: Delfia
Ranieri *et al.*	No	Yes	Cycle	<5 follicles 15 mm	Not applicable	Immunometric: Nichols Inst.
Hall *et al.*	No	No	Patient	Not stated	Clincical	RIA
Bassil *et al.*	No	No	Cycle	Not stated	Clinical	Not stated
Jinno *et al.*	Yes	No	Cycle	Not stated	Not stated	Enzyme immunoassay: Abbott
Bancsi *et al.*	No	Yes	Cycle	Not stated	Ongoing	Immunoan./immunometric: Chiron
Chae *et al.*	Yes	Yes	Cycle	Not stated	Clinical	IRMA: Jeil Japan
Penarrubia *et al.*	Yes	Yes	Cycle	<3 follicles 14 mm	Not applicable	Immunoenzymometric: Technicon
Creus *et al.*	Yes	Yes	Cycle	<3 follicles 14 mm	Not applicable	Immunoenzymometric: Technicon
Fabregues *et al.*	Yes	Yes	Cycle	<3 follicles 14 mm	Not applicable	IRMA: Immunotech
Mikkelsen *et al.*	Yes	No	Retrieval	Not stated	Clinical	Immuno I; Bayer
Van de Stege *et al.*	Yes	Yes	Cycle	<3 follicles 18 mm	Clinical	RIA: Elecsys
Nahum *et al.*	Yes	No	Cycle	<3 follicles 18 mm	Clinical	MEIA: Abbott
Esposito *et al.*	No	Yes	Cycle	Not stated	Live birth	Immuno I; Bayer
Fiçicioğlu *et al.*	Yes	Yes	Cycle	<2 follicles or <5 oocytes	Not applicable	ELISA: Serotec Ltd, UK
Chuang *et al.*	No	Yes	Cycle	Not stated	Ongoing	Chemilum. Immunoassay: Immulite
Yanushpolsky *et al.*	Yes	No	Retrieval	Not stated	Delivery	Techn. Imm. Syst.: Bayer
Erdem *et al.*	Yes	No	Cycle	<5 oocytes or <3 follicles 18 mm	Not applicable	Immunometric: Immulite 2000
Akande *et al.*	Yes	Yes	Cycle	<3 follicles 18 mm	Not applicable	Immunofluorimetric: DELFIA
Kwee *et al.*	Yes	Yes	Cycle	Poor response <6 oocytes	Not applicable	Immunomet.: Amerlite/Delfia

Table III. Performance of basal FSH in the prediction of poor response in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal (= lower than the threshold) FSH result

Author	Cycles (*n*)	FSH threshold value (IU/l)	Prediction of poor response				Pre-FSH probability (%)	Post-FSH probability (%)	Proportion of patients/cycles with abnormal FSH (%)
			Sensitivity	Specificity	LR+	DOR			
Toner *et al*.	1478	10	0.72	0.40	1.2	1.6	7	10	61
		15	0.45	0.75	1.8	2.4	7	15	27
		20	0.31	0.90	3.1	3.9	7	19	12
		25	0.22	0.96	5.5	6.7	7	29	5
Ebrahim *et al*.	111	11.5	0.80	0.93	11.4	49.0	5	38	11
Chan *et al*.	144	4.5	0.94	0.33	1.4	8.2	13	17	71
		6	0.72	0.71	2.5	6.3	13	27	35
Fanchin *et al*.	52	11	0.86	0.45	1.6	4.9	27	37	63
Smotrich *et al*.	292	15	0.00	0.95	0	2.8	2	0	4
Pruksanonda *et al*.	36	4	1.00	0.26	1.4	0.7	3	4	75
		8	1.00	0.71	3.5	4.7	3	10	31
Balasch *et al*.	120	NS	0.50	0.81	2.6	4.3	33	56	29
Gurgan *et al*.	637	10	0.47	0.82	2.6	2.9	16	33	23
		13	0.37	0.92	4.6	4.2	16	47	12
		15	0.33	0.95	6.6	4.9	16	56	9
		20	0.11	0.99	11.0	4.4	16	66	3
Sharif *et al*.	344	5.4	0.91	0.12	1.0	1.3	9	9	89
		10.8	0.31	0.93	4.4	5.9	9	31	9
Evers *et al*.	231	17	0.26	0.97	8.7	10.5	20	69	8
Ranieri *et al*.	177	9.5	0.81	0.65	2.3	8.2	27	48	47
Penarrubia *et al*.	80	Pmodel > 50%	0.83	0.73	3.1	4.5	25	52	41
Creus *et al*.	120	9.45	0.65	0.81	3.4	11.0	33	67	35
Fabregues *et al*.	80	Pmodel > 50%	0.28	0.91	3.1	3.8	35	62	16
Van der Stege *et al*.	87	10	0.60	0.85	4.1	8.8	6	20	17
Nahum *et al*.	272	10	0.22	0.93	3.2	3.8	14	33	9
Fiçicioğlu *et al*.	58	7	0.76	0.76	3.1	9.9	43	70	47
Chuang *et al*.	1045	10	0.32	0.87	2.4	3.1	9	19	15
Erdem *et al*.	32	logistic model	0.63	0.81	3.3	9.7	50	77	41
Akande *et al*.	536	6	0.88	0.50	1.7	6.9	6	10	53
		9	0.59	0.87	4.5	9.7	6	22	16
		12	0.47	0.96	11.3	20.3	6	42	7
Kwee *et al*.	110	4	1.00	0.05	1.1	1.9	26	27	96
		6	0.93	0.40	1.5	8.8	26	36	69
		8	0.72	0.78	3.3	9.2	26	54	35
		10	0.34	0.96	9.3	8.97	26	77	12
		12	0.24	1.00	21.4	28.5	26	89	8

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result; NS, not specified.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table IV. Performance of basal FSH in the prediction of non-pregnancy in IVF patients and shift from pre-test to post-test probability of pregnancy for patients with an abnormal (= lower than the threshold) FSH result

Author	Cycles (*n*)	FSH threshold value (IU/l)	Prediction of non-pregnancy				Pre-FSH probability (%)	Post-FSH probability (%)	Proportion of patients/cycles with abnormal FSH (%)
			Sensitivity	Specificity	LR+	DOR			
Scott *et al.*	758	5	0.85	0.20	1.1	1.5	86	87	85
		10	0.65	0.53	1.4	1.97	86	90	62
		15	0.31	0.84	1.9	2.4	86	92	29
		25	0.08	0.98	4.6	4.9	86	96	7
Padilla *et al.*	91	15	0.40	0.69	1.3	1.5	68	73	37
		20	0.23	0.90	2.3	2.5	68	83	19
Toner *et al.*	1478	10	0.61	0.43	1.1	1.2	83	84	60
		15	0.29	0.89	2.6	1.9	83	93	25
		20	0.13	0.95	2.6	2.5	83	93	10
		25	0.07	1.00	12.0	16.5	83	98	4
Khalifa *et al.*	1110	10	0.58	0.44	1.0	1.1	83	84	58
		15	0.28	0.82	4.4	1.7	83	88	26
		20	0.08	0.93	13.6	1.1	83	84	9
		25	0.06	1.00	11.9	12.6	83	98	5
Ebrahim *et al.*	111	11.5	0.12	0.94	2.0	2.1	85	92	11
Chan *et al.*	144	4.5	0.73	0.54	1.6	3.2	90	94	71
		6	0.37	0.87	2.8	3.9	90	96	35
Huyser *et al.*	139	11.7	0.16	0.96	4.0	4.3	83	95	14
Licciardi *et al.*	452	17	0.19	0.91	2.1	2.3	81	90	17
Smotrich *et al.*	292	15	0.07	1.00	7.6	8.1	65	93	4
Martin *et al.*	1868	20	0.03	1.00	10.1	10.4	84	98	3
Pruksanonda *et al.*	36	4	0.78	0.50	1.6	3.6	89	93	75
		8	0.34	1.00	2.1	2.7	89	94	31
Csemiczky *et al.*	53	7	0.26	1.00	6.8	8.6	58	90	15
Gurgan *et al.*	637	10	0.24	0.80	1.2	1.2	81	84	23
		13	0.14	0.95	2.8	3.1	81	92	12
		15	0.11	0.97	4.3	4.6	81	95	8
		20	0.03	1.00	4.4	4.5	81	95	3
Sharif *et al.*	344	10.8	0.12	0.97	4.0	4.6	70	90	9
Chang *et al.*	149	10	0.13	0.97	4.3	5.5	74	92	10
Evers *et al.*	231	17	0.09	1.00	3.2	3.4	86	95	8
Hall *et al.*	110	9.4	0.77	0.27	1.1	1.95	39	40	75
		11.2	0.60	0.57	1.4	2.0	39	47	50
		13.3	0.33	0.81	1.7	2.0	39	52	25
Bassil *et al.*	83	10	0.45	0.10	0.5	0.1	92	85	49
		15	0.32	0.50	0.6	0.5	92	88	34
		20	0.09	0.80	0.5	0.4	92	85	10
		25	0.04	0.90	0.4	0.4	92	83	5
		30	0.03	1.00	0.5	0.5	92	83	3
Jinno *et al.*	271	15	0.05	0.96	1.1	1.1	65	67	4
Bancsi *et al.*	435	15	0.06	1.00	3.9	4.0	86	96	5
Chae *et al.*	118	8.5	0.46	0.85	3.0	4.6	89	96	42
Mikkelsen *et al.*	130	15	0.34	0.73	1.3	1.4	88	91	33
Van der Stege *et al.*	87	10	0.18	0.85	1.2	1.2	70	73	17
Nahum *et al.*	272	10	0.11	0.96	2.7	2.9	65	83	9
Esposito *et al.*	293	10	0.19	0.91	2.1	2.3	74	85	16
		11.4	0.11	1.00	8.9	9.9	74	96	8
Chuang *et al.*	1045	10	0.18	0.91	2.0	2.2	70	82	15
Yanushpolsky *et al.*	483	10	0.22	0.88	1.9	2.1	62	75	18

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table V. The occurrence of the basal FSH results within a specified likelihoodratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%

Prediction of poor response (pre-test probability = 20%)			Prediction of non-pregnancy (pre-test probability = 80%)		
LR range	Occurrence of test results within this range (%)	Post-test probability poor response (%)	LR range	Occurrence of test results within in this range (%)	Post–test probability non-pregnancy (%)
0–1	68	<20	0–1	63	<80
1–2	15	20–33	1–2	22	80–89
2–3	8	33–43	2–3	9	89–93
3–4	3	43–50	3–4	1	93–94
4–5	2	50–56	4–5	1	94–95
5–6	1	56–60	5–6	1	95–96
6–7	1	60–64	6–7	1	96–96.5
7–8	1	64–67	7–8	1	96.5–97
>8	1	>67	>8	1	>97

Table VI. Characteristics of included studies on AMH (computerized search using the test-specific keywords *anti-mullerian hormone or mullerian inhibiting factor or mullerian inhibiting substance*)

Author	Consecutive	One cycle per couple	Data per	Definition		AMH-assay
				Poor response/Cancel	pregnancy	
Van Rooij *et al.*	Yes	Yes	Cycle	<4 oocytes or <3 follicles	ongoing	Immuno-enzymometric (immunotech-Coulter)
Muttukrishna *et al.*	No	Yes	Cycle	<4 follicles 15 mm.	not applicable	Immuno-enzymometric (immunotech-Coulter)

Table VII. Performance of AMH in the prediction of poor response in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal AMH result

Author	Cycles (*n*)	AMH threshold value (μg/l)	Prediction of poor response				Pre-AMH probability (%)	Post-AMH probability (%)	Proportion of patients/cycles with abnormal AMH (%)
			Sensitivity	Specificity	LR+	DOR			
Van Rooij *et al.*	119	<0.1	0.49	0.94	8.2	14,9	29	77	18
		<0.2	0.54	0.90	5.7	11,3	29	70	23
		<0.3	0.60	0.89	5.6	12,5	29	70	25
Muttukrishna *et al.*	69	<0.1	0.76	0.88	6.6	24.9	25	68	28

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table VIII. Performance of AMH in the prediction of non-pregnancy in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal AMH result

Author	Cycles (*n*)	AMH threshold value (μg/l)	Prediction of non-pregnancy				Pre-AMH probability (%)	Post-AMH probability (%)	Proportion of patients/cycles with abnormal AMH (%)
			Sensitivity	Specificity	LR+	DOR			
Van Rooij *et al.*	119	<0.1	0.22	0.89	1.9	2.2	75	85	19
		<0.2	0.27	0.85	1.8	2.1	75	84	24
		<0.3	0.28	0.81	1.5	1.7	75	81	25

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table IX. Characteristics of included studies on inhibin B (computerized search using the test-specific keyword *inhibin B*)

Author	Consecutive	One cycle per couple	Data per	Definition		Inhibin B assay
				Poor response/Cancel	Pregnancy	
Balasch *et al.*	Yes	Yes	Cycle	<3 follicles 14 mm.	Not applicable	Immunoenzymometric assay (Medgenix)
Seifer *et al.*	Yes	No	Patient	<4 follicles 15 mm.	clinical	ELISA (Serotec Lim. UK)
Hall *et al.*	No	No	Cycle	Not stated	clinical	ELISA (Serotec)
Creus *et al.*	Yes	Yes	Cycle	<3 follicles 14 mm.	Not applicable	Enzyme-linked immunosorbent (Serotec)
Fabregues *et al.*	Yes	Yes	Cycle	<3 follicles 14 mm.	Not applicable	Immunoenzymatic (Medgenix)
Penarrubia *et al.*	Yes	Yes	Cycle	<3 follicles 14 mm.	Not applicable	Immunoenzymometric (Immuno 1; Bayer)
Bancsi *et al.*	Yes	Yes	Cycle	<4 oocytes or <3 follicles 18 mm.	ongoing	Immuno-enzymometric (Serotec)
Fiçicioğlu *et al.*	No	Yes	Cycle	<5 oocytes	Not applicable	ELISA (Serotec)
Erdem *et al.*	Yes	No	Cycle	<5 oocytes (MII) or <3 follicles	Not applicable	Immunosorbent (Serotec)

Table X. Performance of inhibin B in the prediction of poor response in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal inhibin B result

Author	Cycles (*n*)	Inhibin B threshold value (pg/ml)	Prediction of poor response				Pre-inhibin B probability (%)	Post-inhibin B probability (%)	Proportion of patients/ cycles with abnormal inhibin B (%)
			Sensitivity	Specificity	LR+	DOR			
Balasch *et al.*	120	logistic model	0.52	0.80	2.6	4.4	33	57	31
Seifer *et al.*	178	<45	0.53	0.79	2.6	4.3	8	19	24
Creus *et al.*	120	logistic model	0.70	0.63	1.9	3.9	33	48	48
Fabregues *et al.*	80	logistic model	0.32	0.83	1.9	2.3	35	50	23
Penarrubia *et al.*	80	logistic model	0.89	0.29	1.3	3.6	25	30	76
Bancsi *et al.*	120	<45	0.33	0.95	6.9	10	30	75	13
		<53.8	0.39	0.94	6.5	10.1	30	74	16
Fiçicioğlu *et al.*	58	<56	0.81	0.81	4.4	18.0	43	77	45
Erdem *et al.*	32	logistic model	0.69	063	1.8	3.7	50	65	53

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result; NS, not stated.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XI. Performance of the inhibin B in the prediction of non-pregnancy in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal inhibin B result

Author	Cycles (*n*)	Inhibin B threshold value (pg/ml)	Prediction of non-pregnancy				Pre-inhibin B probability (%)	Post-inhibin B probability (%)	Proportion of patients/ cycles with abnormal inhibin B (%)
			Sensitivity	Specificity	LR+	DOR			
Seifer *et al.*	178	<45	0.28	0.92	3.5	4.5	79	93	24
Hall *et al.*	111	<53.8	0.23	0.74	0.9	0.8	39	36	25
		<76.5	0.60	0.56	1.4	1.9	39	46	50
		<105.3	0.77	0.25	1.0	1.1	39	39	76
Bancsi *et al.*	120	<45	0.17	1.00	5.2	6.1	78	94	13
		<53.8	0.19	0.96	5.2	6.2	78	95	16

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XII. The occurrence of the inhibin B results within a specified likelihoodratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%

Prediction of poor response (pre-test probability = 20%)			Prediction of non-pregnancy (pre-test probability = 80%)		
LR range	Occurrence of test results in this range (%)	Post-test probability of poor response (%)	LR range	Occurrence of test results in this range (%)	Post-test probability of non-pregnancy (%)
0–1	60	<20	0–1	79	<80
1–2	22	20–33	1–2	13	80–89
2–3	10	33–43	2–3	4	89–93
3–4	7.8	43–50	3–4	2	93–94
4–5	0.2	50–56	4–5	1	94–95
5–6	0	56–60	5–6	1	95–96
6–7	0	60–64	6–7	0	96–96.5
7–8	0	64–67	7–8	0	96.5–97
>8	0	>67	>8	0	>97

Table XIII. Characteristics of included studies on basal estradiol (computerized search using the test-specific keyword estradiol)

Author	Consecutive	One cycle per couple	Data per	Definition		Estradiol-assay
				Poor response/Cancel	Pregnancy	
Smotrich *et al*.	No	No	Cycle	<2 follicles 16 mm.	Clinical	RIA (Diag. Prod. USA)
Licciardi *et al*.	No	Not stated	Retrieval	Not stated	Ongoing	RIA (Pantax South Monica, CA)
Ranieri *et al*.	No	Yes	Cycle	<5 follicles 15 mm.	Not stated	RIA (Amersham Int. UK)
Evers *et al*.	Yes	Yes	Cycle	<4 follicles 15 mm.	Clinical	RIA (Diag. Prod. USA)
Vazquez *et al*.					Clinical	
Hall *et al*.	No	No	Patient	Not stated	Clinical	Enzyme immunoassay (Abott Lab. USA)
Frattarelli *et al*.	Yes	Yes	Cycle	<3 follicles	Clinical	Immunolite immunoassay (Diag. Pord. USA)
Phophong *et al*.	Yes	Yes	Cycle	<3 follicles 15 mm.	Clinical	RIA (Amersham Int. UK)
Penarrubia *et al*.	Yes	Yes	Cycle	<3 follicles 14 mm.	Not stated	Immunoenzymometric (Immuno I; Bayer)
Mikkelsen *et al*.	Yes	No	Retrieval	Not stated	Clinical	Autoanalyser (Immuno I; Bayer Denmark)
Bancsi *et al*.	Yes	Yes	Cycle	<4 oocytes or <3 follicles 18 mm.	Ongoing	AxSYM immunoanalyser (Abott Lab USA)

Table XIV. Performance of basal estradiol in the prediction of poor response in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal estradiol result

Author	Cycles (*n*)	Estradiol threshold value (pmol/l)	Prediction of poor response				Pre- estradiol probability (%)	Post- estradiol probability (%)	Proportion of patients cycles with abnormal estradiol (%)
			Sensitivity	Specificity	LR+	DOR			
Smotrich *et al*	292	>294	0.83	0.92	10.8	60.0	2	19	9
		>367	0.83	0.97	23.8	138.0	2	33	5
Ranieri *et al*.	177	>350	0.79	0.81	4.1	15.8	27	60	36
Evers *et al*.	213	>220	0.26	0.96	6.5	8.5	16	56	8
Vazquez *et al*.	248	>92	0.64	0.38	1.0	1.1	9	9	62
		>184	0.27	0.71	0.9	0.9	9	8	29
		>275	0.09	0.88	0.7	0.7	9	7	12
		>367	0.05	0.94	0.7	0.7	9	7	6
Frattarelli *et al*.	2476	>73	0.76	0.13	0.9	0.5	14	12	86
		>147	0.34	0.56	0.8	0.7	14	11	43
		>220	0.14	0.88	1.1	1.2	14	15	13
		>294	0.06	0.97	1.95	2.0	14	24	4
		>367	0.03	0.98	2.2	2.2	14	26	2
Phophong *et al*.	305	>250	0.12	0.86	0.8	0.8	9	7	14
Penarrubia *et al*.	80	logistic model	0.70	0.32	1.0	1.1	25	25	69
Bancsi *et al*.	120	>200	0.31	0.74	1.2	1.2	30	33	28
		>250	0.22	0.92	2.7	3.1	30	53	13

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XV. Performance of the basal estradiol in the prediction of non-pregnancy in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal Estradiol result

Author	Cycles (*n*)	Estradiol threshold value (IU/l)	Prediction of non-pregnancy				Pre- estradiol probability (%)	Post- estradiol probability (%)	Proportion of patients/ cycles with abnormal estradiol (%)
			Sensitivity	Specificity	LR+	DOR			
Smotrich *et al.*	292	>294	0.12	0.96	3.1	3.4	65	85	9
		>367	0.08	1.00	8.7	9.4	65	94	5
Licciardi *et al.*	452	>110	0.76	0.37	1.2	1.9	81	84	73
		>165	0.42	0.69	1.3	1.6	81	85	40
		>220	0.20	0.87	1.6	1.8	81	87	19
		>275	0.08	1.00	7.4	8.0	81	97	7
Evers *et al.*	213	>220	0.09	1.00	3.2	3.4	85	94	8
Vazquez *et al.*	248	>92	0.60	0.33	0.9	0.8	70	67	62
		>184	0.29	0.72	1.0	1.0	70	70	29
		>275	0.11	0.85	0.7	0.7	70	63	12
		>367	0.05	0.91	0.5	0.5	70	53	6
Hall *et al.*	120	>108	0.71	0.25	0.95	0.8	38	36	73
		>136	0.47	0.49	0.92	0.9	38	36	49
		>167	0.20	0.72	0.7	0.6	38	30	25
Frattarelli *et al.*	2476	>73	0.84	0.12	0.96	0.8	54	53	86
		>147	0.41	0.55	0.9	0.9	54	52	43
		>220	0.12	0.87	1.0	0.99	54	54	13
		>294	0.04	0.97	1.3	1.3	54	61	4
		>367	0.02	0.99	1.9	1.95	54	70	2
Phopong *et al.*	305	>250	0.13	0.83	0.8	0.8	77	72	14
Mikkelsen *et al.*	132	>200	0.22	1.00	3.9	4.7	89	96	20
Bancsi *et al.*	120	>200	0.27	0.70	0.9	0.9	78	76	28
		>250	0.12	0.85	0.8	0.8	78	73	13

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XVI. The occurrence of the basal estradiol results within a specified likelihoodratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%

Prediction of poor response (pre-test probability = 20%)			Prediction of non-pregnancy (pre-test probability = 80%)		
LR range	Occurrence of test results in this range (%)	Post-test probability of poor response (%)	LR range	Occurrence of test results in this range (%)	Post-test probability of non-pregnancy (%)
0–1	83	<20	0–1	82	<80
1–2	12	20–33	1–2	17	80–89
2–3	3	33–43	2–3	1	89–93
3–4	1	43–50	3–4	0	93–94
4–5	1	50–56	4–5	0	94–95
5–6	0	56–60	5–6	0	95–96
6–7	0	60–64	6–7	0	96–96.5
7–8	0	64–67	7–8	0	96.5–97
>8	0	>67	>8	0	>97

Table XVII. Characteristics of included studies on basal AFC (computerized search using test-specific keywords *antral follicle count or antral follicle number*)

Author	Consecutive	One cycle per couple	Data per	Definition		Diameter follicles (mm)	ultrasonograph
				Poor response/Cancel	Pregnancy		
Chang *et al*.	Yes	No	Cycle	<2 follicles 18 mm.	Ongoing	2–5	Accuson 120XP/10: 7 MHz probe
Ng *et al*.	Yes	Yes	Cycle	<3 follicles 15 mm.	Clinical	Not stated	Aloka SSD-620: 5 MHz probe
Frattarelli *et al*. (2000)	No	Yes	Cycle	<3 follicles	Not stated	2–10	Acuson 128: 7 MHz probe
Sharara *et al*.	Yes	No	Cycle	Not stated	Clinical	2–8	Not stated
Hsieh *et al*.	Yes	No	Cycle	No oocytes or poor follicle growth	Clinical	2–10	Acuson Aspen: 4 MHz probe
Nahum *et al*.	Yes	No	Cycle	<3 follicles 18 mm.	Clinical	2–6	General electric RT-X200: 6.5 MHz probe
Fisch *et al*.	Yes	Yes	Cycle	Not stated	Clinical	Not stated	Not stated
Bancsi *et al*.	Yes	Yes	Cycle	<4 oocytes or <3 follicles 18 mm.	Clinical/ongoing	2–5	Toshiba Capasee SSA-220A: 7.5 MHz probe
Frattarelli *et al*. (2003)	Yes	Yes	Cycle	<3 follicles	Not stated	2–10	Acuson 128: 7 MHz probe
Järvelä *et al*.	Yes	Yes	Cycle	<4 follicles	Clinical	2–5	Kretz Combison 530D
Kupesic *et al*.	Yes	Yes	Cycle	Not stated	Clinical	Not stated	Combison 530D: 7.5 MHz probe
Yong *et al*.	No	Yes	Cycle	<4 oocytes or cancel	Clinical	2–10	Toshiba Eccocee: 7 MHz probe
Fiçicioğlu *et al*.	Yes	Yes	Cycle	<5 oocytes	Not stated	≥2	General Electric Alfa Logic 200: 5 MHZ probe
Erdem *et al*.	Yes	Yes	Cycle	<3 follicles 14 mm. or <5 oocytes (MII)	Not stated	Not stated	Aloka SSD-1000: 5 MHz probe
Durmusoglu *et al*.	No	No	Cycle	Poor follicle growth or <3 oocytes (MII)	Not stated	2–10	GE Logiq200: 6.5 MHz probe

Table XVIII. Performance of basal AFC in the prediction of poor response in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal AFC result

Author	Cycles (*n*)	AFC threshold value (*n*)	Prediction of poor response				Pre-AFC probability (%)	Post-AFC probability (%)	Proportion of patients/cycles with abnormal AFC (%)
			Sensitivity	Specificity	LR+	DOR			
Chang *et al*.	149	<3	0.73	0.96	19.7	65	10	69	11
Ng *et al*.	128	<4	0.33	0.92	4.2	5.7	2	9	9
		<6	0.80	0.76	3.3	13	2	11	27
		<9	0.80	0.40	1.3	2.7	2	5	61
Frattarelli *et al*. (2000)	278	<10	0.87	0.41	1.5	4.7	8	12	61
Sharara *et al*.	127	<4	0.53	0.73	1.9	3.0	15	26	31
Hsieh *et al*.	372	<3	0.61	0.94	10.0	23	5	34	9
Nahum *et al*.	272	<6	0.95	0.69	3.1	42	14	33	39
Bancsi *et al*.	120	<4	0.61	0.88	5.1	12	30	69	27
		<6	0.81	0.77	3.6	14	30	60	40
Frattarelli *et al*. (2003)	267	<4	0.30	0.96	7.4	10	9	41	6
Järvelä *et al*.	45	<4	0.86	0.84	5.4	32	16	50	27
Yong *et al*.	47	<4	0.09	0.97	3.3	3.2	23	50	4
		<6	0.36	0.89	3.3	4.6	23	50	17
Fiçicioğlu *et al*.	58	<7	0.77	0.41	1.3	2.3	43	50	66
Erdem *et al*.	32	logistic model	0.75	0.63	2.0	5.1	50	67	56
Durmusoglu *et al*.	91	<6.5	0.85	0.74	3.3	16	26	53	41

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XIX. Performance of basal AFC in the prediction of non-pregnancy in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal AFC result

Author	Cycles (*n*)	AFC threshold value (*n*)	Prediction of non-pregnancy				Pre-AFC probability (%)	Post-AFC probability (%)	Proportion of patients/cycles with abnormal AFC (%)
			Sensitivity	Specificity	LR+	DOR			
Chang *et al*.	149	<3	0.13	0.96	3.6	3.6	83	94	11
Ng *et al*.	128	<4	0.07	0.83	0.4	0.4	86	73	9
		<6	0.26	0.78	1.2	1.2	86	88	26
		<9	0.60	0.33	0.9	0.7	86	61	85
Sharara *et al*.	127	<4	0.27	0.64	0.8	0.7	56	49	31
Hsieh *et al*.	372	<3	0.12	0.98	6.9	6.7	68	94	9
Nahum *et al*.	272	<6	0.54	0.87	4.0	7.9	64	88	39
Fisch *et al*.	200	<10	0.24	0.89	2.2	2.6	59	76	19
Bancsi *et al*.	107	<4	0.34	0.88	2.9	3.8	68	86	27
		<6	0.45	0.68	1.4	1.7	68	75	41
Järvelä *et al*.	45	<4	0.26	0.71	0.9	0.9	69	67	27
Kupesic *et al*.	56	<4	0.33	0.96	8.3	11.8	61	92	22
Yong *et al*.	47	<4	0.08	0.92	0.9	1.0	76	75	9
		<6	0.16	0.90	1.6	1.7	79	86	27

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XX. The occurrence of the AFC results within a specified likelihood ratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%

Prediction of poor response (pre-test probability = 20%)			Prediction of non-pregnancy (pre-test probability = 80%)		
LR range	Occurrence of test results in this range (%)	Post-test probability of poor response (%)	LR range	Occurrence of test results in this range (%)	Post-test probability of non-pregnancy (%)
0–1	68	<20	0–1	77	<80
1–2	10	20–33	1–2	16	80–89
2–3	4	33–43	2–3	5	89–93
3–4	6	43–50	3–4	0	93–94
4–5	0	50–56	4–5	2	94–95
5–6	0	56–60	5–6	0	95–96
6–7	0	60–64	6–7	0	96–96.5
7–8	0	64–67	7–8	0	96.5–97
>8	12	>67	>8	0	>97

Table XXI. Characteristics of included studies on ovarian volume (OVVOL) (computerized search using the test-specific keyword *ovarian volume*)

Author	Consecutive	One cycle per couple	Data per	Definition		Ovarian volume (ml) definition	Ultrasonography equipment
				Poor response/Cancel	Pregnancy		
Syrop *et al.*	Yes	Yes	Cycle	<2 follicles 18 mm.	Clinical	Total <8.6 ml, smallest <3 ml	General Elect. 3600: 5 MHz probe
Lass *et al.*	Yes	Yes	Cycle	<3 follicles 17 mm.	Clinical	MOV <3 ml	Kretz Comb. 410: 5–7.5 Mhz probe
Sharara *et al.*	Yes	Yes	Cycle	Poor follicle development	Not stated	MOV <3 ml	Performa: 6.5 MHz probe
Schild *et al.*	Yes	Yes	Cycle	Not stated	Biochemical	MOV <3 ml	Voluson 530D : 7.5 MHz probe
Bancsi *et al.*	Yes	Yes	Cycle	<4 oocytes or <3 follicles 18 mm	Clinical	Total <7 ml or <8.6 ml	Toshiba SSA: 7.5 MHZ probe
Kupesic *et al.*	Yes	Yes	Cycle	Not stated	Biochemical	Total <7 ml	Combison 530 D: 7.5 Mhz probe
Järvelä *et al.*	Yes	Yes	Cycle	<4 follicles	Clinical	MOV <7 ml or <3 ml	Kretz Comb 530
Fiçicioğlu *et al.*	Yes	Yes	Cycle	<5 oocytes	Not stated	Total <4.9 ml	General Electric Alfa Logic 200: 5 MHz probe
Erdem *et al.*	Yes	Yes	Cycle	<3 follicles 14 mm. or <5 oocytes (MII)	Clinical	MOV < 2.98 ml	Aloka SSD1000: 5 MHz probe
Frattarelli *et al.*	Yes	Yes	Cycle	<3 follicles	Biochemical	MOV < 2 ml or < 3 ml	Acuson 128: 7 MHz probe

MOV, mean ovarian volume.

Table XXII. Performance of the ovarian volume in the prediction of poor response in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal volume result

Author	Cycles (*n*)	Volume threshold value (ml)	Prediction of poor response				Pre-volume probability (%)	Post-volume probability (%)	Proportion of patients/cycles with abnormal volume (%)
			Sensitivity	Specificity	LR+	DOR			
Syrop *et al.*	188	<8.6	0.25	0.86	1.78	2.0	13	21	15
		<3	0.17	0.91	1.95	2.1	13	22	10
Lass *et al.*	140	<3	0.45	0.93	6.75	11.5	14	53	12
Sharara *et al.*	73	<3	0.80	0.72	2.86	10.3	7	17	32
Schild *et al.*	152	<3	0.11	0.90	1.10	1.1	18	20	10
Bancsi *et al.*	120	<8.6	0.61	0.73	2.23	4.2	30	49	38
		<7	0.39	0.85	2.51	3.5	30	52	23
Kupesic *et al.*	56	<7	0.86	0.87	6.49	39.4	12	46	22
Järvelä *et al.*	60	<3	0.08	0.94	1.30	1.3	18	25	6
		<7	0.55	0.67	1.67	2.5	18	27	37
Fiçicioğlu *et al.*	58	<4.9	0.73	0.53	1.50	2.7	43	53	59
Erdem *et al.*	32	<2.98	0.75	0.81	4.00	13.0	50	80	47
Frattarelli *et al.*	267	<2	0.17	0.94	2.83	3.2	9	21	7
		<3	0.35	0.82	1.89	1.4	9	15	20

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XXIII. Performance of the ovarian volume in the prediction of non-pregnancy in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal volume result

Author	Cycles (*n*)	Volume threshold value (ml)	Prediction of non-pregnancy				Pre-volume probability (%)	Post-volume probability (%)	Proportion of patients/cycles with abnormal volume (%)
			Sensitivity	Specificity	LR+	DOR			
Syrop *et al.*	188	<8.6	0.17	0.87	1.23	1.3	65	69	15
		<3	0.11	0.93	1.44	1.5	65	72	10
Lass *et al.*	140	<3	0.12	0.88	0.97	0.96	89	88	12
Schild *et al.*	152	<3	0.12	0.97	3.60	3.9	80	93	10
Bancsi *et al.*	120	<8.6	0.47	0.71	1.58	2.1	68	77	41
		<7	0.27	0.79	1.33	1.5	68	74	25
Kupesic *et al.*	56	<7	0.33	0.96	8.00	11.5	60	92	22
Järvelä *et al.*	60	<3	0.08	0.96	1.80	1.9	63	75	6
		<7	0.42	0.73	1.54	1.9	63	73	37
Erdem *et al.*	32	<2.98	0.70	0.92	8.40	25.7	63	93	47
Frattarelli *et al.*	267	<2	0.10	0.96	2.46	2.6	47	68	7
		<3	0.22	0.82	1.27	1.4	47	53	20

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XXIV. The occurrence of the ovarian volume test results within a specified likelihood ratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%

Prediction of poor response (pre-test probability = 20%)			Prediction of non-pregnancy (pre-test probability = 80%)		
LR range	Occurrence of test results in this range (%)	Post-test probability of poor response (%)	LR range	Occurrence of test results in this range (%)	Post-test probability of non-pregnancy (%)
0–1	54	<20	0–1	68	<80
1–2	16	20–33	1–2	31	80–89
2–3	30	33–43	2–3	3	89–93
3–4	0	43–50	3–4	0	93–94
4–5	0	50–56	4–5	0	94–95
5–6	0	56–60	5–6	0	95–96
6–7	0	60–64	6–7	0	96–96.5
7–8	0	64–67	7–8	0	96.5–97
>8	0	>67	>8	0	>97

Table XXV. Characteristics of included studies on ovarian stromal flow (OSF) (computerized search using the test-specific keyword *ovarian ovarian stromal blood flow*)

Author	Consecutive	One cycle per couple	Data per	Definition		OSF parameter	Ultrasonography equipment
				Poor response/Cancel	Pregnancy		
Kupesic *et al.*	Yes	Yes	Cycle	Not applicable	Biochemical	Peak systolic velocity	Combison 530 D, 7.5 Mhz probe

Table XXVI. Performance of ovarian stromal flow (OSF) in the prediction of non-pregnancy in IVF patients and shift from pre-test to post-test probability of pregnancy for patients with an abnormal (= higher than the threshold) ovarian stromal flow result

Author	Cycles (*n*)	OSF threshold value (flow index)	Prediction of non-pregnancy				Pre-OSF probability (%)	Post-OSF probability (%)	Proportion of patients/cycles with abnormal OSF (%)
			Sensitivity	Specificity	LR+	DOR			
Kupesic *et al*	56	<11	0.31	0.96	7.7	4.1	60	92	20
		≤13	0.85	0.23	1.1	1.5	60	64	82

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XXVII. Characteristics of included studies on the CCCT (computerized search using the test-specific keyword clomiphene citrate challenge test)

Author	Consecutive	One cycle per couple	Data per	Definition		FSH-assay
				Poor response/Cancel	Pregnancy	
Tanbo *et al*.	No	Yes	Cycle	Cancel <3 follicles	Not stated	RIA: Amerlex
Tanbo *et al*.	No	No	Cycle	Cancel <2 follicles	Clinical	Fluoroimmunoassay: Delfia
Loumaye *et al*.	No	Yes	Cycle	Cancel <2 follicles 20 mm	Not stated	Immunoradiometr.: IRMA
Tanbo *et al*.	No	No	Cycle	Cancel <3 follicles	Ongoing	Fluoroimm.assay: Delfia
Csemiczky *et al*.	No	No	Cycle	Not stated	Clinical	RIA: Diagn Prod. Inc.
Kahraman *et al*.	No	Yes	Cycle	Not stated	Ongoing	Immunometr.: Diagn. Prod. Corp.
Vd Stege *et al*.	Yes	Yes	Cycle	Cancel <3 follicles 18 mm	Clinical	RIA: Roche Diagn.
Csemiczky *et al*.	No	Yes	Cycle	Cancel <3 follicles 17 mm	Ongoing	RIA: Farmos Group
Yanushpolsky *et al*.	Yes	No	Retrieval	Not stated	Delivery	Techn. Imm. Syst.: Bayer Corp.
Kwee *et al*.	Yes	Yes	Cycle	Poor response <6 oocytes	Not stated	Immunomet.: Amerlite/Delfia
Erdem *et al*.	Yes	Yes	Cycle	Cancel <4 follicles 15 mm or Poor response <5 oocytes	Clinical	Chemolum. Immunometr. Assay
Hendriks *et al*.	Yes	Yes	Cycle	Poor response <4 oocytes or cancel no follicle growth	Ongoing	AxSYM immunoanal.: Abbott Lab.

Table XXVIII. Performance of the CCCT in the prediction of poor response in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal CCCT result

Author	Cycles (*n*)	FSH threshold value (IU/L)	Prediction of poor response				Pre-CCCT probability (%)	Post-CCCT probability (%)	Proportion of patients/cycles with abnormal CCCT (%)
			Sensitivity	Specificity	LR+	DOR			
Tanbo *et al*.	109	Day 7 > 26	0.55	0.97	17.7	37.8	40	92	24
Tanbo *et al*.	70	Day 10 > 26	0.75	0.47	1.4	2.7	46	55	63
Loumaye *et al*.	114	Day 3 + 10 > 26.03	0.83	0.86	6.0	31.0	5	25	18
Tanbo *et al*.	165	Day 10 > 12	0.57	0.91	5.9	12.5	49	85	33
Kahraman *et al*.	198	Day 10 > 10	0.43	0.76	1.8	2.4	25	37	29
Vd Stege *et al*.	51	Day 3 or 10 > 10	0.50	0.82	2.7	4.4	4	10	20
Csemiczky *et al*.	279	Day 10 > 10	0.54	0.84	3.3	6.1	25	53	26
Kwee *et al*.	56	Day 3 + 10 > 14	0.93	0.68	2.9	30.1	27	52	48
		Day 3 + 10 > 16	0.80	0.83	4.7	19.4	27	63	34
		Day 3 + 10 > 18	0.73	0.95	15.0	53.6	27	85	23
		Day 3 + 10 > 20	0.60	0.98	24.6	60.0	27	90	18
		Day 3 + 10 > 22	0.53	0.98	21.9	45.7	27	89	16
Erdem *et al*.	32	Day 3 or 10 > 10	0.69	0.88	5.5	15.4	50	85	41
Hendriks *et al*.	63	Day 10 > 10	0.65	0.87	5.0	12.2	27	65	27
		Day 10 > 15	0.35	0.96	8.1	12.0	27	75	13

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XXIX. Performance of the CCCT in the prediction of non-pregnancy in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal CCCT result

Author	Cycles (*n*)	FSH threshold value (IU/L)	Prediction of non-pregnancy				Pre-CCCT probability (%)	Post-CCCT probability (%)	Proportion of patients/cycles with abnormal CCCT (%)
			Sensitivity	Specificity	LR+	DOR			
Tanbo *et al.*	70	Day 10 > 26	0.66	0.80	3.3	7.8	93	98	63
Loumaye *et al.*	114	Day 3 + 10 > 26.03	0.23	0.96	6.5	8.2	76	95	19
Tanbo *et al.*	165	Day 10 > 12	0.34	0.86	2.4	3.1	96	98	33
Csemiczky *et al.*	53	Day 10 > 7	0.61	0.96	14.5	35.4	58	95	37
Kahraman *et al.*	198	Day 10 > 10	0.30	0.85	2.4	3.0	92	96	29
Vd Stege *et al.*	51	Day 3 or 10 > 10	0.22	0.84	1.4	1.5	63	70	20
Csemiczky *et al.*	140	Day 10 > 10	0.30	0.97	8.6	11.8	79	97	24
Yanushpolsky *et al.*	483	Day 10 > 10	0.36	0.82	2.0	2.5	62	76	29
Erdem *et al.*	32	Day 3 or 10 > 10	0.45	0.67	1.4	1.6	63	69	41
Hendriks *et al.*	63	Day 10 > 10	0.27	0.73	1.0	1.0	76	77	27
		Day 10 > 15	0.13	0.87	0.9	0.9	76	75	13

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XXX. Characteristics of included studies on the EFORT (computerized search using the test-specific keyword *EFORT*)

Author	Consecutive	One cycle per couple	Data per	Definition		Estradiol-assay
				Poor response/Cancel	Pregnancy	
Fanchin *et al.*	Yes	Yes	Cycle	<3 oocytes	Not stated	Estradiol-60 Amerlite (Kodak clin. Diagn. UK)
Kwee *et al.*	Yes	Yes	Cycle	<6 oocytes	Not stated	Amerlite (Amersham UK)
Yong *et al.*	No	Yes	Cycle	<4 oocytes or cancel	Not stated	Radioimmunoassay

Table XXXI. Performance of the EFORT in the prediction of poor response in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal EFORT result

Author	Cycles (*n*)	Estradiol threshold value (pmol/l)	Prediction of poor response				Pre-EFORT probability (%)	Post-EFORT probability (%)	Proportion of patients/cycles with abnormal EFORT (%)
			Sensitivity	Specificity	LR+	DOR			
Fanchin *et al.*	52	< 110	0.79	0.92	2.7	42.8	27	79	27
Kwee *et al.*	54	< 110	0.64	0.68	1.98	3.7	26	41	41
		< 120	0.64	0.65	1.8	3.3	26	39	43
		< 130	0.71	0.65	2.0	4.6	26	42	44
		< 140	0.79	0.60	1.96	5.5	26	41	50
		< 150	0.86	0.58	2.0	8.1	26	41	54
Yong *et al.*	46	< 124	0.50	0.68	1.6	2.2	17	25	35

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XXXII. Characteristics of included studies on GAST (computerized search using the test-specific keyword *gonadotrophin releasing hormone agonist stimulation test*)

Author	Consecutive	One cycle per couple	Data per	Definition		Estradiol assay
				Poor response/Cancel	Pregnancy	
Padilla *et al.*	No	No	Cycle	Not stated	Clinical	RIA (Diagnostic Products USA)
Winslow *et al.*	Yes	Yes	Cycle	Not stated	Clinical	Radioimmunoassay (Pantex CA)
Ranieri *et al.*	No	Yes	Cycle	<5 follicles 15 mm	Not stated	RIA (Amersham Int. UK)
Hendriks *et al.*	Yes	Yes	Cycle	<4 oocytes or <3 follicles 18 mm	Ongoing	AxSYM immunoanalyser (Abbott Lab USA)

Table XXXIII. Performance of GAST in the prediction of poor response in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal GAST result

Author	Cycles (*n*)	Estradiol threshold value (pmol/l)	Prediction of poor response				Pre-GAST probability (%)	Post-GAST probability (%)	Proportion of patients/cycles with abnormal GAST (%)
			Sensitivity	Specificity	LR+	DOR			
Winslow *et al*.	228	$E_2/E_1 < 2$	0.58	0.95	11.5	26.1	5	39	8
Ranieri *et al*.	177	$\Delta E_2 < 180$	0.89	0.86	6.4	53.0	27	70	34
Hendriks *et al*.	57	$\Delta E_2 < 80$	0.32	0.97	12.0	17.1	33	86	12
		$\Delta E_2 < 100$	0.37	0.89	3.5	4.6	33	64	19
		$\Delta E_2 < 180$	0.68	0.79	3.3	8.1	33	62	37

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XXXIV. Performance of GAST in the prediction of non-pregnancy in IVF patients and shift from pre-test to post-test probability of non-pregnancy for patients with an abnormal GAST result

Author	Cycles (*n*)	Estradiol threshold value (pmol/l)	Prediction of non-pregnancy				Pre-GAST probability (%)	Post-GAST probability (%)	Proportion of patients/cycles with abnormal GAST (%)
			Sensitivity	Specificity	LR+	DOR			
Padilla *et al*.	97	$E_2/E_1 < 2$	0.27	0.90	2.8	3.5	68	86	22
Winslow *et al*.	228	$\Delta E_2 < 50$	0.42	0.70	1.4	1.69	77	82	39
		$\Delta E_2 < 75$	0.66	0.53	1.4	2.2	77	82	62
		$\Delta E_2 < 100$	0.76	0.38	1.2	1.92	77	80	73
Hendriks *et al*.	57	$\Delta E_2 < 80$	0.16	1.00	2.4	2.7	79	89	12
		$\Delta E_2 < 100$	0.24	1.00	3.6	4.6	79	92	19
		$\Delta E_2 < 180$	0.40	0.75	1.6	2.0	79	86	37

DOR, diagnostic odds ratio; LR+, likelihood ratio for a positive test result.
If a study reported on multiple threshold values, data for all threshold values are shown.

Table XXXV. The occurrence of the GAST volume results within a specified likelihood ratio (LR) range and the concomitant post-test probabilities of poor response and non-pregnancy, given a prevalence of poor response of 20% and non-pregnancy of 80%

Prediction of poor response (pre-test probability = 20%)			Prediction of non-pregnancy (pre-test probability = 80%)		
LR range	Occurrence of test results in this range (%)	Post-test probability of poor response (%)	LR range	Occurrence of test results in this range (%)	Post-test probability of non-pregnancy (%)
0–1	31	<20	0–1	70	<80
1–2	8	20–33	1–2	22	80–89
2–3	5	33–43	2–3	2	89–93
3–4	6	43–50	3–4	6	93–94
4–5	3	50–56	4–5	0	94–95
5–6	4	56–60	5–6	0	95–96
6–7	5	60–64	6–7	0	96–96.5
7–8	7	64–67	7–8	0	96.5–97
>8	30	>67	>8	0	>97

Table XXXVI. Characteristics of included studies on multi-variate models (computerized search using the test-specific keywords *multifactor, multivariate, prediction model and logistic model*)

Author	Consecutive	One cycle per couple	Data per	Definition	
				Poor response/Cancel	Pregnancy
Balasch *et al.*	Yes	Yes	Cycle	<2 follicles 17 mm. or <5 follicles 14 mm	Not applicable
Fabregues *et al.*	Yes	Yes	Cycle	<3 follicles 14 mm	Not applicable
Ranieri *et al.*	No	Yes	Cycle	<5 follicles 15 mm	Not applicable
Creus *et al.*	Yes	Yes	Cycle	<3 follicles 14 mm	Not applicable
Bancsi *et al.*	Yes	Yes	Cycle	<4 oocytes or <3 follicles 18 mm	Not applicable
Van Rooij *et al.*	Yes	Yes	Cycle	<4 oocytes or <3 follicles	Not applicable
Muttukrishna *et al.*	No	Yes	Cycle	<4 follicles 15 mm	Not applicable
Erdem *et al.*	Yes	Yes	Cycle	Cancel <4 follicles 15mm or poor response <5 oocytes	Not applicable
Durmusoglu *et al.*	No	No	Cycle	Poor follicles growth or <3 oocytes (MII)	Not applicable

Table XXXVII. Performance of multi-variate models in the prediction of poor response in IVF patients and shift from pre-test to post-test probability of poor response for patients with an abnormal test result

Author	Cycles (*n*)	Test Model	Prediction of poor response				Pre-test probability (%)	Post-test probability (%)	Proportion of patients/cycles with abnormal test (%)
			Sensitivity	Specificity	LR+	DOR			
Balasch *et al.*	120	Age + FSH	0.53	0.81	2.8	4.8	33	58	30
		Age + inhibin B	0.59	0.67	1.8	2.8	33	48	42
		Inhibin B + FSH	0.57	0.69	1.8	3.0	33	48	40
		Age + FSH + inhibin B	0.39	0.89	3.5	5.2	33	64	21
Fabregues *et al.*	80	FSH + Inhibin B	0.42	0.86	3.0	4.4	35	63	24
Ranieri *et al.*	177	FSH + GAST	0.97	0.55	2.2	39.5	33	45	59
Creus *et al.*	120	Age + FSH	0.83	0.77	3.6	16.3	33	65	43
		Age + inhibin B	0.74	0.50	1.5	2.8	33	43	58
		FSH + inhibin B	0.77	0.73	2.9	9.1	33	58	44
		Age + FSH + inhibin B	0.83	0.77	3.6	16.3	33	65	43
Bancsi *et al.*	120	FSH + inhibin B	0.58	0.94	9.7	21.6	30	81	22
		AFC + inhibin B	0.69	0.88	5.6	16.3	30	71	29
		AFC + FSH	0.72	0.93	10.3	34.2	30	81	27
		AFC + inihbin B + FSH	0.75	0.95	15	57.0	30	87	26
Van Rooij *et al.*	119	AMH + inhibin B + FSH	0.69	0.91	7.7	22.5	29	75	27
Muttukrishna *et al.*	69	FSH + inhibin B + AMH	0.63	0.83	3.7	8.3	25	65	29
Erdem *et al.*	32	CCCT + age	0.81	0.69	2.6	9.5	50	72	56
		CCCT + age + OVVOL + AFC	0.81	0.75	3.2	12.8	50	76	53
Durmusoglu *et al.*	91	Age + AFC	0.52	0.88	4.3	7.9	26	62	23

AFC, antral follicle count; CCCT, clomiphene citrate challenge test; DOR, diagnostic odds ratio; GAST, gonadotrophin agonist stimulation test; LR+, likelihood ratio for a positive test result; OVVOL, ovarian volume.
If a study reported on multiple threshold values, data for all threshold values are shown.