A systematic review of tests predicting ovarian reserve and IVF outcome

J.Kwee1, D.J.Hendriks2, B.W.Mol3, C.B.Lambalk1,4 and F.J.Broekmans2

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

4To 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-induced 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 because 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
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,
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 hyper-stimulation 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 accurate 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. 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) See Supplementary Data.

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 χ²-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
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.[...], 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 et al., 2002; Chuang et al., 2003; Fiçicio 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 I. 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 II See Supplementary Data. Sensitivity and specificity points, as plotted in Figure 4, were heterogeneous between studies ($\chi^2$-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 judge-ment 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 III See Supplementary Data. Again, sensitivity and specificity points plotted in Figure 5 were heterogeneous between studies ($\chi^2$-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 IV (See Supplementary Data) 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 II 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%...
based on a pre-test rate of 20%, again occur only in a very limited number of patients (Table IV). 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 Supplementary Table II.

**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 V (See Supplementary Data) and in Figure 6 Homogeneity could not be rejected for sensitivity and specificity (χ²-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.

**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 Supplementary Table III. Variation among the definitions

---

**Estimated ROC curve**

Sensitivities and specificities for the prediction of non-pregnancy by AMH, as calculated from each study, are summarized in Table VI. 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).

---

**AMH - Poor Response**

![Figure 6](http://humupd.oxfordjournals.org/)

**AMH - Non Pregnancy**

![Figure 7](http://humupd.oxfordjournals.org/)
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

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 Supplementary Data. 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 ($\chi^2$-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 VIII. Sensitivity and specificity as plotted in Figure 9 were heterogeneous among studies ($\chi^2$-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 LR-s 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 IX 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.

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 Supplementary Table IV. 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 X. 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 ($\chi^2$-test statistic: $P$-value for sensitivity <0.001 and $P$-value for specificity 0.002). The Spearman correlation coefficient for sensitivity and specificity was –0.94. 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 XI. Again, sensitivity and specificity as plotted in Figure 11 were heterogeneous between studies ($\chi^2$-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.

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 XII 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 McClanock, 2000; Hsieh et al., 2001; Nahum et al., 2001; Bancsi et al., 2002a; Erdem et al., 2002; Fisch and Sher, 2002; Ficicioglu 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 Supplementary Table V. 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

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.
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 XIII. 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 ($\chi^2$-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 XIV. Again, sensitivity and specificity as plotted in Figure 13 were heterogeneous between studies ($\chi^2$-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 XV 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 $\sim8$, 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 $\sim8$ 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.](image)

For prediction of non-pregnancy, the extremely low AFC that is necessary to obtain a moderate positive likelihood ratio of $\sim5$, 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 XV). 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 $\sim1$).

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 $\sim5$% 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; Füçicioğlu 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 Supplementary Table VI. 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.,
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 XVI. Homogeneity for both sensitivity and specificity had to be rejected ($\chi^2$-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 XVII. As with the data for ovarian response, homogeneity for sensitivity had to be rejected. However, specificity appeared to be homogeneous ($\chi^2$-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.

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 XVIII 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 XVIII 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 Supplementary Table VII and XIX). 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 Supplementary Table VIII. 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 XX. Homogeneity could not be rejected for sensitivity (χ²-test statistic: P-value 0.09), but had to be rejected for specificity (χ²-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 XXI. Homogeneity was rejected for both sensitivity and specificity (χ²-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).](http://humupd.oxfordjournals.org/)

Page 12 of 19
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 fulfill 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 Supplementary Table IX.

Accuracy of poor response prediction

The individual values for sensitivity and specificity pairs are summarized in Table XXII 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 fulfills 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 Supplementary Table X. 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 XXIII. 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 ($\chi^2$-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 XXIV. Again, sensitivity and specificity, as shown in Figure 20, were heterogeneous between studies ($\chi^2$-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 XXV 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 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 Supplementary Table XI. 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
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).

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 fulfill 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 outcome as 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.

Supplementary material

Supplementary data is available online at http://humupd.oxfordjournals.org/.

References


Chang MY, Chiang CH, Chiu TH, 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.

Downloaded from http://humupd.oxfordjournals.org/ at Pennsylvania State University on February 26, 2014
Chuang CC, Chen CD, Chao KH, Chen SU, Ho HH 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.


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.


Review of tests predicting ovarian reserve


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 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 18,34–44.


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


