Value of ovarian reserve testing before IVF: a clinical decision analysis

Ben W.Mol1,2,3,4,9, Tamara E.M.Verhagen5, Dave J.Hendriks6, John A.Collins7,8, Arri Coomarasamy3, Brent C.Opmeer2 and Frank J.Broekmans6

1Department of Obstetrics and Gynecology, Centre for Reproductive Medicine, Academic Medical Center, Amsterdam, 2Department of Obstetrics and Gynecology, Máxima Medical Centre, Veldhoven, The Netherlands, 3Department of Obstetrics and Gynaecology, Birmingham Women’s Hospital, Birmingham, UK, 4Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, Amsterdam, 5Department of Obstetrics and Gynaecology, Academic Hospital Maastricht, Maastricht, 6Department of Reproductive Medicine, University Medical Center Utrecht, Utrecht, The Netherlands, 7Department of Obstetrics and Gynaecology, McMaster University, Hamilton, Ontario and 8Department of Obstetrics and Gynaecology, Dalhousie University, Halifax, Nova Scotia, Canada

9To whom correspondence should be addressed at: Department of Obstetrics and Gynecology, Academic Medical Center, P.O.Box 22660, 1100 DD Amsterdam, The Netherlands. E-mail: b.mol1@chello.nl

BACKGROUND: To assess the value of testing for ovarian reserve prior to a first cycle IVF incorporating patient and doctor valuation of mismatches between test results and treatment outcome. METHODS: A decision model was developed for couples who were considering participation in an IVF programme. Three strategies were evaluated: (I) withholding IVF without prior testing, (II) testing for ovarian reserve, and then deciding on IVF treatment if ovarian reserve was estimated to be sufficient, and (III) treatment with IVF without prior ovarian reserve testing. The outcome considered was the birth of a child. The valuation of the combination of the strategy conducted and the outcome accomplished was expressed on a distress scale in units of ‘IVF cycles that were performed in vain’. Correct treatment with IVF and correct withholding of IVF were considered to bring no distress. The distress of withholding IVF in case pregnancy occurred is consequently specified by the ratio of the expected distress after incorrect withholding IVF to the expected distress after incorrect performing IVF (distress ratio). We interviewed both patients and doctors to determine realistic estimates for this distress ratio. RESULTS: The value of testing for ovarian reserve depends strongly on the expected pregnancy rate after IVF as well as on the valuation of the incorrect decisions from testing. For realistic ranges of the success rate after IVF and for distress ranges as were measured, treatment of all couples without testing was found to generate less distress than testing for ovarian reserve. The sensitivity and specificity of testing for ovarian reserve has to improve to 50 and 96% respectively, to make testing a valuable strategy. CONCLUSION: Based on the decision analysis, where current test accuracy and preference inventory among patients and physicians were used, testing for ovarian reserve seems not useful for current IVF programmes.

Key words: artificial reproductive treatment/decision analysis/ovarian reserve testing

Introduction

The ability of the ovaries to respond to exogenous gonadotrophin stimulation and to develop several follicles with subsequent embryo selection after IVF of retrieved oocytes is one of the success factors in IVF. The occurrence of a poor response, often with cancellation of the treatment cycle, is a common problem. Identification of potential poor responders before the actual start of the IVF treatment would potentially enable physicians to counsel these patients on the prospects of the treatment that they are about to begin and even to actively withhold treatment in certain cases.

We have previously reported on the accuracy of several tests to assess the ovarian reserve prior to IVF treatment. Among these tests, the Antral Follicle Count (AFC) has been shown to be a significantly better predictor of poor response than basal FSH (Hendriks et al., 2005). However, all available tests to date perform inaccurately in the prediction of pregnancy.

The clinical question of whether ovarian reserve testing before IVF does more good than harm depends not only on their discriminative accuracy, but also on the consequences coupled to the test result and the valuation of the test based consequences. Net benefit and harm of a strategy that incorporates ovarian reserve testing before IVF need to be compared to other clinically relevant strategies. An attractive strategy, which is explicitly or implicitly used in current clinical practice, is the use of the actual response in the first stimulation cycle. A poor response could imply discontinuation of treatment, as the chance of IVF success would be approaching zero.
Recent work has shown that only the combination of an abnormal ovarian reserve test and a poor ovarian response in the first cycle (the expected poor response case) indicates very low prospects in subsequent cycles (Klinkert et al., 2004).

Although currently available tests have only very limited predictive capacity for pregnancy, a combination of existing tests in multivariable models or the introduction of new tests could resolve this issue. However, the requirements of such composite or new tests in terms of diagnostic accuracy in order to be useful in clinical practice are unknown. In this paper, we use a decision analytic framework, modelling the test accuracy as well as patients and doctors perception of these outcomes. Value judgments of both incorrect start of IVF (i.e. IVF without a pregnancy) and incorrect withholding of IVF (i.e. not starting IVF where pregnancy would have occurred) are inventoried among a group of patients and doctors working in the field of infertility. In a decision analysis we evaluated whether the expected benefits based on probabilities and value judgements of outcomes would justify testing for ovarian reserve or treatment without prior testing.

Materials and methods

General problem definition and analytical approach

We used a representation of the decision problem in terms of three available options, analogous to the approach described by Pauker and Kassirer (1980): (I) withholding IVF treatment, (II) test for ovarian reserve and then decide for IVF treatment in case ovarian reserve was judged sufficient, and (III) treatment of the couple without prior testing for ovarian reserve and withholding further treatment in case of poor response. The expected value of an alternative depends on the probability of possible outcomes when applying the strategy and the subjective value that can be attached to each of the four possible outcomes can be expressed in terms of the expected distress that combinations, incorrect starting or withholding of treatment indeed will generate distress. The central issue here concerns the distress of ‘Incorrect no IVF’ relative to ‘Incorrect IVF’. In other words, how much more is the couple assumed to suffer from incorrect withholding IVF as compared to incorrect starting of IVF. This ratio of the distress of ‘Incorrect No IVF’ and the distress of ‘Incorrect IVF’ is referred to as the ‘distress ratio’ (Thornton and Lilford 1995; Van der Meulen et al., 2004).

Treatment strategies

We compared the three mutually exclusive alternatives: (I) withhold treatment without testing for ovarian reserve, (II) test for ovarian reserve, and then decide for IVF treatment in case ovarian reserve was judged to be sufficient, and (III) treatment of the couple without prior testing for ovarian reserve and withholding further treatment in case of poor response (Figure 1).

Outcomes

The foremost relevant clinical outcome to evaluate is the extent to which different strategies result in ongoing pregnancies. However, we hypothesize that from the perspective of the couple, a pregnancy that fails to occur after IVF is valued differently from a couple not becoming pregnant without IVF. Four possible outcomes are (a) correct start of IVF (i.e. start of an IVF cycle resulting in pregnancy), (b) incorrect start of IVF (i.e. start of an IVF cycle not resulting in pregnancy), (c) incorrect refraining from IVF (i.e. withholding an IVF cycle that would have resulted in pregnancy) and (d) correct refraining from IVF (i.e. withholding an IVF cycle that would not have resulted in pregnancy) (Figure 2). These outcomes will be further referred to as ‘Correct IVF’, ‘Incorrect IVF’, ‘Incorrect No IVF’ and ‘Correct No IVF’ respectively. In order to appropriately evaluate different strategies, the subjective valuation of these outcomes of the test and treatment cycle and the probabilities that they occur have to be taken into account.

Figure 1. Decision tree representing the three strategies, the probabilities of different outcomes and the values [in terms of distress (DIS = distress)] attached to these outcomes.
Table 1. Baseline parameter values, parameter ranges used in sensitivity analysis and references used

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Variable</th>
<th>Baseline</th>
<th>Range</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of non-pregnancy</td>
<td>PROB_pregnant</td>
<td>0.80</td>
<td>0.40–1.00</td>
<td>Hendriks et al., 2005</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>Sens</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>Spec</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distress ratio</td>
<td>DIS_treat_notpregnant*</td>
<td>100</td>
<td>10–250</td>
<td>Hendriks et al., 2005</td>
</tr>
</tbody>
</table>

*DIS_treat_notpregnant is set at 1.

For the inventory of preferences among physicians working in the field of reproductive medicine, we approached gynaecologists and fertility doctors in six other clinics performing IVF in the Netherlands, and presented them a situation in which a fictive test was done prior to IVF that would allow or not a couple starting IVF treatment. Again, they were asked how much worse they would valuate the incorrect withholding of IVF based on this fictive test as compared to the incorrect start of IVF based on the same test.

Decision analyses

The following parameters, based on a review of the literature, are used in the decision analyses (Table I). The probability of non-pregnancy in a first cycle IVF was set at 80%, sensitivity and specificity at 24 and 90% respectively, and the distress ratio was set at 100, i.e. the incorrect withholding of IVF was judged 100 times worse than the incorrect starting of IVF. The expected distress for each of the strategies was then calculated by weighing the disutilities associated with each outcome with the probability of that outcome (see Appendix).

Subsequently, we calculated the expected distress for a wide range of prevalences of non-pregnancy (40–100%) and a wide range of distress ratios (10–250). A distress ratio of 5 was excluded because we gave more importance to the patients’ distress than the doctors’ distress on their behalf, and 1000 was excluded because it was ridiculously high. We then plotted the expected distress as a function of the prevalence of non-pregnancy. From these graphs, we were able to derive the strategy that would result in the lowest expected distress.

The accuracy required to make testing a better strategy than treatment of all couples is evaluated by assessing the threshold for the prevalence of non-pregnancy at which testing for ovarian reserve becomes superior to treatment of all couples. In other words, we should focus on the threshold for the prevalence of non-pregnancy at which treating without testing results in an expected distress that is equal to that of testing with subsequent treatment of those with expected normal ovarian reserve (Pauker and Kassirer 1980). If the prevalence of non-pregnancy is lower than this threshold, treatment of all couples should be the preferred treatment, whereas otherwise a strategy starting with testing for ovarian reserve would be preferable.

We calculated the expected distress for each of the three strategies for each combination of sensitivity and specificity, when prevalence of non-pregnancy and distress-ratio are fixed. The strategies I, treat nobody without testing, and III, treat all without testing, result in a fixed expected distress that is independent of sensitivity and specificity. For strategy II, test all couples for ovarian reserve, and then decide on treatment, the expected distress will be minimal in case the test characteristics sensitivity and specificity are perfect (i.e. both 100%). This distress increases once test properties decrease. The threshold is then determined by that combination of sensitivity and specificity at which the expected distress of strategy II, testing for ovarian reserve, is equal to the lowest of the two other strategies. This threshold has been plotted into a Receiver-Operating-Characteristic (ROC)-space for different distress ratios for a probability of non-pregnancy of 80 and 50%.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Variable</th>
<th>Baseline</th>
<th>Range</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of non-pregnancy</td>
<td>PROB_pregnant</td>
<td>0.80</td>
<td>0.40–1.00</td>
<td>Hendriks et al., 2005</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>Sens</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>Spec</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distress ratio</td>
<td>DIS_treat_notpregnant*</td>
<td>100</td>
<td>10–250</td>
<td>Hendriks et al., 2005</td>
</tr>
</tbody>
</table>

*DIS_treat_notpregnant is set at 1.

Inventory of preferences among patients and doctors

To make an inventory of the preferences, we performed face-to-face interviews among subfertile couples that were scheduled for IVF in the Utrecht Medical Centre in Utrecht, The Netherlands. If indicated, IVF is offered without restrictions to couples in which the female’s age is <41 years. In the interview, couples received a written explanation of the use of a fictive prediction test prior to IVF designed to decide whether couples would be allowed or not to enter the treatment. Also, the imperfections of testing were discussed and written information was further explained by one of the authors (DH). Subsequently, they were asked how much worse they would valuate the incorrect withholding of IVF based on this fictive test as compared to the incorrect start of IVF based on the same test.
Results

Inventory of preferences among patients and doctors

We performed an inventory among patient couples that were indicated to initiate IVF treatment. For these couples a structured interview was designed, including a description of the use and characteristics of a given, fictive, ovarian reserve test, and the question to value the incorrect withholding of IVF as compared to the incorrect start of IVF. In total 25 couples were asked to participate of which 20 (80%) agreed and finished the structured interview. The valuations by couples resulted in distress ratios ranging from 50 to 1000, with a median of 250, indicating that couples in general expected 250 times more distress by incorrect refusal of IVF than starting IVF without getting pregnant.

For the 36 clinicians working in the field of assisted reproduction, we prepared a written survey in which they were asked to value incorrect withholding of IVF as compared to an incorrect start of IVF. Surveys were then sent to representatives of all 7 IVF centres in the Netherlands (n = 36). The response rate was 70%. From the 25 responding clinicians, one felt unable to answer the questions posed. From the answers of the other 24 the distress ratios ranged between 5 and 250, with a median of 25 (Figure 3). The difference between the distress ratios recorded among physicians and patient couples was statistically significant (P-value <01, Mann–Whitney U-test). The results obtained in this inventory were used as baseline and range parameters in the subsequent decision analyses.

Decision analyses

In the baseline analysis, the expected distress for the three strategies are 20, 2.6 and 0.80 ‘incorrect start of IVF equivalents’ for strategies I, II and III respectively, indicating that treatment of couples without prior testing (strategy III) generates the least distress. In subsequent analyses, the probability of non-pregnancy was varied between 40 and 100%, and sensitivity and specificity remained constant at their baseline values (24 and 90% respectively), corresponding with the results of a previous meta-analysis on the accuracy of AFC (Hendriks et al., 2005). Figure 4A shows the expected distress for the three strategies, when the distress ratio is presumed to be 100, i.e. the incorrect withholding of IVF is valued 100 times as distressing as the incorrect performance of IVF. When the probability of non-pregnancy was below 77% (i.e. an IVF success rate of 23% per cycle or higher), strategy III had the lowest expected distress. In case of a probability of non-pregnancy between 97.5 and 99.5% (i.e. IVF success rates of 2.5–5% per cycle), the strategy with testing for ovarian reserve (strategy II) was expected to have the lowest expected distress, whereas above a probability of non-pregnancy of 99.5% (IVF success rate per cycle lower than .5%), no treatment (strategy I) would bring the lowest expected distress.

Figure 4B, C and D depict the situation for distress ratios of 10, 50 and 250 respectively, showing that with the incorrect withholding of IVF distress increases much more rapidly with increasing distress ratio compared to incorrect application of IVF treatment. These figures show the same pattern as for a distress ratio of 10, although the ‘treat all-to-test all’ threshold and the ‘test all-to-no treatment’ threshold both shift to higher probabilities of failure of IVF. The probability of IVF failure at which testing brings less distress than the other two strategies lies between 77 and 93% for a distress ratio of 10, 94 and 98.5% for a distress ratio of 50, and between 99 and 99.7% for a distress ratio of 250.

Subsequently we addressed the question to which degree test accuracy should be improved to make testing a better strategy than treatment of all couples without testing by determining those levels of sensitivity and specificity at which testing for ovarian reserve will become superior to treatment of all couples without testing. Instead of keeping sensitivity and specificity constant at their baseline values, prevalence of non-pregnancy after IVF was fixed at 80% and the distress ratio was set at 10 (i.e. not starting IVF in a situation where pregnancy would have occurred would be valued 10 times as distressing as starting IVF in a situation where pregnancy would not have occurred). Figure 5A shows an ROC curve for this situation, indicating that testing would be superior over treatment of all couples when the accuracy of ovarian reserve testing is better than the line between the left lower corner (0% sensitivity and 100% specificity) and the combination of 100% sensitivity and 60% specificity. The point at which sensitivity is 0% and specificity is 100% implies that at this point none of the cases of diminished ovarian reserve are detected, and that all couples are therefore treated. This test strategy has therefore at this combination of sensitivity and specificity an expected distress that is equal to the immediate treatment of all couples (–0.8 ‘incorrect start of IVF equivalent’). Once specificity subsequently decreases, the sensitivity has to increase in order to keep the expected distress equal to –0.8 ‘incorrect start of IVF equivalent’, which is the expected distress of treatment of all couples. For a sensitivity of 100%, this is the case at a specificity of 60%.

As can also be seen from this ROC diagram, only a small part of the summary ROC curve for the AFC lies to the left of
1820

this threshold line. This implies that only with the use of an extreme cut-off for an abnormal test the desired level of expected distress is realized. Figure 5A also shows that with the increase of the distress ratio to 25, 50 and 100, (i.e. the valuation of incorrect withholding of IVF as more distressing), the demands placed upon the test accuracy sharply rise. As a result, for distress ratios of 25 and higher, testing for ovarian reserve is always inferior to treatment of all couples without testing.

Figure 5B shows the situation for a probability of non-pregnancy of 50%. As can be seen from this ROC curve, the demands put upon test accuracy increase in this situation to such an extent that treatment of all couples is always superior to testing, even if the distress ratio is as low as 10.

Discussion
This study shows that the value of ovarian reserve testing prior to IVF for individual couples strongly depends on the prevalence of IVF failure as well as on the valuation of the false positive (incorrect withholding IVF) or false negative (incorrect performing IVF) outcomes. According to the decision model used in this study, testing for ovarian reserve is only worthwhile when the distress ratio does not exceed 10, and the prevalence of non-pregnancy is around 80%. When we performed an inventory of these preferences among physicians in the field of reproductive medicine and IVF-indicated patient couples, patients expected to suffer far less distress from IVF that was incorrectly started than physicians thought they would. Although we did not evaluate test accuracy but rather the valuation of false-positive and false-negative outcomes, these data show that the clinical value of available tests, given their present accuracy level, is low and that the routine use of these tests in clinical practice can be reasonably questioned.

Whether individual differences in rating the true outcome as compared to the outcome predicted by a test can affect the value of that test is only rarely addressed. In the vast majority of test evaluations, predetermined threshold risks are used, above which testing is advised as clinically useful and offered to the patient. As explained in the Methods section, the threshold risk between starting IVF and testing for ovarian reserve on one hand and between testing for ovarian reserve and not starting IVF on the other hand is fully determined by the distress
Ovarian reserve testing pre IVF

1821

1821

ratio (Van der Meulen et al., 1999). At a prevalence of non-pregnancy of 77%, which is in our opinion a rather modest estimate of the success rate of IVF per cycle in present programmes, the distress ratio of 10 resulted in equal expected distress between treatment of all couples and testing for ovarian reserve. The interviews indicated that median preferences of both patients and doctors were far above this distress ratio. This implies that from both the patient and physician point of view not any ovarian reserve test should be used for the decision whether or not to start IVF. Instead, for counselling purposes or for adjustment of treatment schedules, the test may elicit much less disappreciation of false-positive test results and hence be more clinically acceptable and applicable. However, if testing only implies such minor consequences then health improvement or efficiency gain may become too low to justify the burden of testing.

Decision analysis for diagnostic issues is more complex than for therapeutic issues. In therapeutic issues, the outcome measure is usually unidimensional, as the patient does or does not benefit from treatment. In diagnostic issues, the outcome is influenced both by the discriminatory performance of the test under study and by the effects of subsequent treatment as compared to no treatment. Thus, there are four possible outcomes, i.e. true-positive test results, false-positive test results, true-negative test results and false-negative test results. We decided to assign no distress to consequences of true-positive results and true-negative results, since these two outcomes represent the optimal result from a diagnostic point of view. The other two consequences of test results were then valued on a negative scale, and combined in a distress ratio.

Variations in the pregnancy rate of the IVF programme have an important effect on the value of the ovarian reserve tests. If the pregnancy rate increases from 20 to 50%, the test accuracy of ovarian reserve testing has to improve very strongly, as can be seen from the comparison between Figure 5 A and B. As the comparison with the summary ROC curve for the AFC shows, its test accuracy is too low to be of clinical value, even at distress ratios as low as 10.

In the field of reproductive medicine, recently a shift was made from surrogate outcomes such as ovarian response, ovum pick-up rate and embryo-transfer rate towards more relevant outcomes, of which live birth rate is the most important one. In the literature on testing for ovarian reserve, most studies report on ovarian response rate and/or pregnancy rate. Since pregnancy rate is the outcome most relevant for the patient, we used that one in our study.

The number of performed cycles was not specified in the decision analysis. One could speculate that if the observed outcome (pregnancy) would be calculated over a series of IVF cycles, the distress ratio would become decreased, since three cycles of IVF performed without a pregnancy will be valued worse than one cycle of IVF without a pregnancy. On the other hand, test performance may well improve if the predicted outcome may become realized in more than one attempt, as more exposures give better opportunity to separate the reproductive successful cases from those that are bound to fail. Three cycles

Figure 5. (A,B) Receiver operating characteristic curve for prevalences of non-pregnancy of 80% (Figure 3A) and 50% respectively. The straight lines represent the threshold at which an ovarian reserve test becomes accurate enough to become more valuable than treatment of all couples without testing. Combinations of sensitivity and specificity left of these lines represent test combinations that are good enough to justify testing. As can be seen from the figures, a higher distress ratio and a lower prevalence of non-pregnancy make immediate treatment more attractive than testing. The figures also show a summary ROC curve for AFC in the prediction of non-pregnancy, as was estimated in a meta-analysis (Hendriks et al., 2005). This ROC curve is only good enough if the prevalence of non-pregnancy is high, and if the distress ratio is below 10. DR = distress ratio.
of IVF will correspond with an expected pregnancy rate of 50%. As Figure 5B shows, the distress ratios have to be low to make ovarian reserve testing potentially of value. In view of the high distress ratio that we measured, it is unlikely that patients will change these valuations strongly when three cycles are performed instead of one. Only if the performance curve for the test would clearly shift to the left and the valuations would change to such an extent that the threshold lines would fan to the right, testing may become useful.

The technique of using a ratio to assign values to outcomes of a medical procedure is known as magnitude estimation (Stevens, 1971; Froberg and Kane 1989). An argument for its use is that it is very easy to ask patients to quantify how much worse they think that the not performing IVF would be compared with performing IVF in vain. Such a ratio represents the ovarian reserve testing decision as a choice between certain outcomes and therefore does not account for the attitude of patients to risk (Watson and Buede 1987). As an alternative to measuring the values of the outcomes of the IVF programme which incorporates this risk attitude, one could use the more complex standard reference gambling technique, where values are measured through the elicitation of a series of preferences between lotteries. Most people are averse to risk, which means that the distress ratios used in our study might underestimate the expected distress after withholding IVF incorrectly, especially if the distress ratio is relatively high. Empirical research is required to investigate to what extent the use of standard reference gambles would result in different decisions regarding testing for ovarian reserve.

We conclude that the value of the testing for ovarian reserve should also be determined on the basis of individual preferences, and that the use of a predetermined threshold risk above which IVF is withheld should not be the only factor that directs the decision on whether or not to use a certain test. The distress ratio as introduced in this study may be an instrument to incorporate subjective judgement of the outcomes of an IVF programme for individual couples in clinical practice. As our interviews demonstrated, elicitation of these values is possible in clinical practice, since couples answered the question in a more or less consistent way. Moreover, their preferences were clearly different from that of the doctors, thus indicating that doctors should incorporate the preferences of their patients in their decisions, rather than deciding for them. The result of this study supports the concept of formal decision making in individual couples to complement the process of counselling.

References
As the start of treatment in women that will conceive (Exp.\textsubscript{U\textunderscore treat\_pregnant}) is assumed to bring no distress (i.e. Exp\_DIS set a zero), this can be simplified to

(3B) \quad \text{Prob}\textunderscore\text{notpregnant} \times \text{Exp}\textunderscore\text{DIS}\textunderscore\text{treat\_notpregnant}

If the prevalence of not becoming pregnant after IVF-ET is set at 80% and an expected distress of treatment with IVF in case of not becoming pregnant is 1, then the expected distress of this strategy will be −2.6 ‘incorrect start of IVF equivalents’ (shown in Figure 4 as black dot).