Comparison of basal follicle-stimulating hormone versus the clomiphene citrate challenge test for ovarian reserve screening

Tarun Jain, M.D., a,d Michael R. Soules, M.D., b and John A. Collins, M.D. c

Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts; University of Washington, Seattle, Washington; and McMaster University, Hamilton, Ontario, Canada

Objective: To compare the value of basal follicle-stimulating hormone (FSH) measurement vs. the clomiphene citrate challenge test (CCCT) in predicting the ability to achieve a pregnancy in women who are undergoing infertility treatment.

Design: Meta-analysis.

Setting: All studies that evaluated either basal FSH or the CCCT for determining the likelihood of pregnancy.

Patient(s): Infertility population undergoing treatment, which was defined as patients undergoing ovulation induction, IUI, or in vitro fertilization (IVF).

Intervention(s): None.

Main Outcome Measure(s): Diagnostic test characteristics were calculated and pooled using standard methods. Inability to achieve a pregnancy with treatment was considered as the “disease.”

Result(s): Twelve studies on basal FSH (with 6,296 patients, mean age 33.8) and seven studies on the CCCT (with 1,352 patients, mean age 34.5) fit our criteria and were analyzed. For basal FSH and the CCCT, the sensitivities were 6.6% (95% confidence interval [CI] 5.9, 7.3%) and 25.9% (95% CI 23.0, 29.0%), respectively, and specificities were 99.6% (95% CI 99.1, 99.9%) and 98.1% (95% CI 96.5, 99.1%), respectively. For “disease” prevalence ranging from 40%–100%, for basal FSH and the CCCT, the positive predictive values ranged from 91.7%–100% and 90.1%–100%, respectively, and negative predictive values ranged from 61.5%–0.0% and 66.5%–0.0%, respectively.

Conclusion(s): Basal FSH and the CCCT are similar in predicting the ability to achieve a clinical pregnancy in women undergoing infertility treatment. With either test, a normal result is not useful, but an abnormal result virtually confirms that pregnancy will not occur with treatment. (Fertil Steril 2004;82:180–5. ©2004 by American Society for Reproductive Medicine.)

Key Words: Day 3 FSH, CCCT, reproductive aging, infertility, assisted reproductive technology, IVF, meta-analysis

Reproductive age women experience a decline in fecundity as a function of ovarian aging, which correlates with increased chronological age (1). This decline is associated with ovarian follicular depletion and diminished oocyte quality, a process referred to as diminishing ovarian reserve (2–4). These changes are associated with a decrease in inhibin B (but not E2) along with an increase in follicle-stimulating hormone (FSH) (5). Because age alone is not as reliable a predictor of fecundity (4), other methods have been established to evaluate ovarian reserve in patients presenting to infertility clinics. The two most widely used tests for ovarian reserve screening in an infertility population are basal FSH measurement or the clomiphene citrate challenge test (CCCT). The results of these tests are often used to counsel women regarding their reproductive potential and to make treatment decisions. For instance, women with an abnormal test are often counseled to pursue oocyte donation or adoption instead of infertility treatment.

Basal FSH testing requires a single serum measurement of FSH in the early follicular phase of the menstrual cycle on either day 2, 3, or 4 (6, 7). The CCCT, however, requires FSH...
The purpose of our study was to compile the published data on the use of basal FSH and the CCCT for predicting treatment outcome (pregnancy) in an infertility population, and to calculate and compare the value of these tests.

**MATERIALS AND METHODS**

We completed electronic searches of MEDLINE (January 1966 to March 2003), SCIENCE Direct Database (January 1966 to March 2003), the Database of Abstracts of Reviews of Effectiveness (January 1966 to March 2003), and the Cochrane Database of Systematic Reviews (January 1991 to March 2003). Various combinations of the following search terms were used: basal, day 3, follicle-stimulating hormone, FSH, clomiphene citrate challenge test, CCCT, predictive value, and ovarian reserve. The bibliographies of included studies were also hand searched for any additional relevant articles. Only published data were included in our analysis, and there was no language restriction. Two of the authors (Tarun Jain and John A. Collins) performed independent searches to avoid missing any relevant studies.

Inclusion and exclusion criteria were determined before data abstraction. We included all published studies that evaluated either basal FSH or the CCCT for predicting treatment outcome in an infertility clinic population. Infertility treatment included ovulation induction, intrauterine insemination, or the use of assisted reproductive technologies (ARTs). The outcome of interest in all studies was achieving a clinical pregnancy, which was defined as observation of at least a gestational sac on ultrasonography. We excluded abstracts, case reports, studies with outcomes other than pregnancy, and studies that did not report sufficient data to permit completion of a $2 \times 2$ contingency table.

Data from all qualifying studies was abstracted into $2 \times 2$ tables and independently analyzed by two authors (Tarun Jain and John A. Collins). Data from each study were pooled into either a basal FSH group or a CCCT group. Diagnostic test properties (i.e., sensitivity, specificity, and likelihood ratios) were calculated using standard methods (12). Failure to achieve a pregnancy with infertility treatment was considered as the “disease.” Assessment of heterogeneity among the studies within the basal FSH group and the CCCT group was performed by plotting the sensitivity and specificity for each group on a receiver operating characteristic (ROC) plot (12). In addition, a weighted average diagnostic odds ratio and the associated heterogeneity were estimated with the use of fixed and random effects models. The estimates were similar and the fixed effects inverse variance model is presented in the Results section. A $P$ value $<.05$ was considered to indicate significant heterogeneity, suggesting that the studies were too different to combine and generate summary statistics.

The positive and negative predictive values for basal FSH and the CCCT were calculated and plotted as a function of varying prevalence of disease (no pregnancy with treatment) (13). Using the pooled sensitivity and specificity for basal

---

**TABLE 1**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Question addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>How good is this test at picking up women who cannot conceive?</td>
</tr>
<tr>
<td>Specificity</td>
<td>How good is this test at correctly excluding women who can conceive?</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>If a woman has an abnormal test, what is the probability that she cannot conceive?</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>If a woman has a normal test, what is the probability that she can conceive?</td>
</tr>
<tr>
<td>Likelihood ratio of a positive test</td>
<td>How much more likely is an abnormal test to be found in a woman who cannot conceive than in a woman who can?</td>
</tr>
<tr>
<td>Likelihood ratio of a negative test</td>
<td>How much more likely is a normal test to be found in a woman who can conceive than in a woman who cannot?</td>
</tr>
</tbody>
</table>


blood testing on cycle day 3 and cycle day 10, as well as orally administering 100 mg of clomiphene citrate on cycle days 5 through 9 (4). An elevated FSH value (based on a predetermined cutoff) on either day 3 (basal FSH or CCCT) or day 10 (CCCT) is “abnormal,” indicating diminished ovarian reserve and a poor response to infertility treatment (8).

Compared with basal FSH, the CCCT is more than twice the cost and is associated with greater inconvenience and potential side effects for patients. Besides the moderate expense of 10 (50 mg) tablets of clomiphene citrate (CC), the CCCT requires a patient to undergo blood testing on two separate days. Furthermore, CC can be associated with numerous and troublesome side effects, such as vasomotor flushes (10.4%), abdomen/pelvic discomfort/distension/bloating (5.5%), nausea/vomiting (2.2%), breast discomfort (2.1%), visual disturbances (1.5%), headache (1.3%), and abnormal uterine bleeding (1.3%) (9).

It has been hypothesized that the CCCT is better than basal FSH for predicting infertility treatment outcome because two levels of FSH are obtained, and the addition of CC creates a “provocative” test that unmasks patients who might not be detected by basal FSH screening alone (10). This hypothesis, however, remains unproven, leaving it unclear whether one test is better than the other as a predictor of treatment outcome in the infertile population. To compare screening tests, one must be aware of the characteristics of these tests (Table 1) (11). To date, we are unaware of any studies that adequately evaluated the characteristics of both tests in the same population of women.
FSH and the CCCT, the following formulas, based on Bayes’ theorem, were used to make the calculations (14):

Positive predictive value =
\[
\frac{(\text{prevalence})(\text{sensitivity})}{(\text{prevalence})(\text{sensitivity}) + (1 - \text{specificity})(1 - \text{prevalence})}
\]

Negative predictive value =
\[
\frac{(1 - \text{prevalence})(\text{specificity})}{(1 - \text{prevalence})(\text{specificity}) + (1 - \text{sensitivity})(\text{prevalence})}
\]

RESULTS

We identified 249 studies that met our initial broad search strategy, of which 27 studies met our inclusion criteria. Four studies (15–18) were excluded because the outcome of interest was not achieving pregnancy, and four studies (19–22) were excluded due to insufficient presentation of data to complete a 2 × 2 table. Of the remaining 19 studies, 12 evaluated basal FSH (23–34) and 7 evaluated the CCCT (4, 8, 35–39) (Table 2). All studies were observational in nature, and no studies were available that calculated all the diagnostic test properties for either basal FSH or the CCCT. The average patient ages were 33.8 and 34.5 in the basal FSH and CCCT groups, respectively. Follicle-stimulating hormone cutoffs varied among the studies due to the use of different FSH assays and reagents. However, the ROC plot showed relatively little scatter between studies (Fig. 1), and the test of heterogeneity for both basal FSH and the CCCT was not significant (P = .756 and P = .926, respectively). Table 3 outlines the pooled results for both basal FSH and the CCCT. A guide to using likelihood ratios in clinical practice suggests that positive likelihood ratios >10 or negative likelihood ratios <0.1 can provide high diagnostic accuracy (40). The false positive rates for basal FSH and the CCCT were 0.079% (5 out of 6,296) and 0.67% (9 out of 1,352), respectively. Positive and negative predictive values (PPV and NPV) for both tests based on varying prevalence of “disease” are plotted in Figure 2.

DISCUSSION

Our study demonstrates that basal FSH and the CCCT have similar value in predicting the ability to achieve a clinical pregnancy in women undergoing infertility treatment. With either test, an abnormal result virtually confirms that pregnancy will not occur with treatment (due to a high PPV), but a normal result is not useful (due to a low NPV).
An ideal diagnostic test has high sensitivity and specificity. With most tests, however, a trade-off generally exists between these two measures of test validity. Any decision regarding specific criteria for acceptable levels of sensitivity and specificity involves weighing the consequences of leaving cases undetected (i.e., false negatives) against erroneously classifying healthy individuals as diseased (i.e., false positives). Sensitivity should be increased at the expense of specificity when the penalty associated with missing a case is high, when the disease is communicable, or when subsequent diagnostic evaluations of positive screening tests are associated with minimal risk and cost (41). For example, in screening for cancer, a test with high sensitivity is critical because you do not want to miss anyone with the disease. To achieve this, it may be acceptable to have a lower specificity and have some false positives. In the case of ovarian reserve screening, the philosophy is the opposite. We want a test with high specificity to minimize falsely counseling patients that they would have no chance of getting pregnant with treatment. The trade-off is to sacrifice sensitivity by letting some patients go through treatment, even though they too may have a poor chance of success. Both basal FSH and the CCCT fit this strategy by having low sensitivity, but high specificity.

Altering the criterion for positivity or abnormality will influence both the sensitivity and specificity of a test. In the case of basal FSH and the CCCT, the FSH cutoff value is the

---

### TABLE 3

Diagnostic test properties estimated from the included studies.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Year</th>
<th>Sensitivity (%) (95% CI)</th>
<th>Specificity (%) (95% CI)</th>
<th>LR+</th>
<th>LR−</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day 3 FSH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bancsi et al.</td>
<td>2000</td>
<td>5.9 (3.7, 8.8)</td>
<td>100 (94.2, 100)</td>
<td>59.0</td>
<td>0.94</td>
</tr>
<tr>
<td>Buyalos et al.</td>
<td>1997</td>
<td>12.8 (6.2, 22.3)</td>
<td>100 (85.2, 100)</td>
<td>128.2</td>
<td>0.87</td>
</tr>
<tr>
<td>Ebrahim et al.</td>
<td>1993</td>
<td>12.5 (6.4, 21.3)</td>
<td>95.7 (78.1, 99.9)</td>
<td>2.9</td>
<td>0.91</td>
</tr>
<tr>
<td>Esposito et al.</td>
<td>2002</td>
<td>11.1 (7.2, 16.0)</td>
<td>100 (95.3, 100)</td>
<td>110.6</td>
<td>0.89</td>
</tr>
<tr>
<td>Evers et al.</td>
<td>1998</td>
<td>9.0 (5.4, 13.9)</td>
<td>100 (89.1, 100)</td>
<td>90.5</td>
<td>0.91</td>
</tr>
<tr>
<td>Gurgan et al.</td>
<td>1997</td>
<td>3.5 (2.1, 5.4)</td>
<td>100 (96.9, 100)</td>
<td>34.7</td>
<td>0.97</td>
</tr>
<tr>
<td>Huys et al.</td>
<td>1995</td>
<td>17.2 (10.3, 26.1)</td>
<td>95.0 (83.1, 99.4)</td>
<td>3.4</td>
<td>0.87</td>
</tr>
<tr>
<td>Martin et al.</td>
<td>1996</td>
<td>3.5 (2.6, 4.6)</td>
<td>100 (98.7, 100)</td>
<td>35.0</td>
<td>0.96</td>
</tr>
<tr>
<td>Pearlstone et al.</td>
<td>1992</td>
<td>22.5 (13.5, 34.0)</td>
<td>100 (76.8, 100)</td>
<td>225.4</td>
<td>0.77</td>
</tr>
<tr>
<td>Scott et al.</td>
<td>1989</td>
<td>8.9 (6.8, 11.5)</td>
<td>98.7 (95.4, 99.8)</td>
<td>6.9</td>
<td>0.92</td>
</tr>
<tr>
<td>Smotrich et al.</td>
<td>1995</td>
<td>6.5 (2.8, 12.4)</td>
<td>100 (96.4, 100)</td>
<td>65.0</td>
<td>0.93</td>
</tr>
<tr>
<td>Toner et al.</td>
<td>1991</td>
<td>7.2 (5.7, 8.9)</td>
<td>100 (99.0, 100)</td>
<td>71.7</td>
<td>0.93</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>6.6 (5.7, 7.2)</td>
<td>99.6 (99.0, 99.9)</td>
<td>17.2</td>
<td>0.94</td>
</tr>
<tr>
<td>CCCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Csemiczky et al.</td>
<td>1996</td>
<td>61.3 (48.2, 78.2)</td>
<td>100 (84.6, 100)</td>
<td>612.9</td>
<td>0.39</td>
</tr>
<tr>
<td>Hofmann et al.</td>
<td>1996</td>
<td>21.2 (15.0, 28.4)</td>
<td>96.8 (89.0, 99.6)</td>
<td>6.7</td>
<td>0.81</td>
</tr>
<tr>
<td>Loumaye et al.</td>
<td>1990</td>
<td>22.7 (14.5, 32.9)</td>
<td>100 (86.8, 100)</td>
<td>227.3</td>
<td>0.77</td>
</tr>
<tr>
<td>Navot et al.</td>
<td>1987</td>
<td>47.2 (30.4, 60.5)</td>
<td>93.3 (68.1, 99.8)</td>
<td>7.1</td>
<td>0.57</td>
</tr>
<tr>
<td>Scott et al.</td>
<td>1993</td>
<td>14.8 (9.4, 21.7)</td>
<td>97.9 (92.5, 99.7)</td>
<td>7.0</td>
<td>0.87</td>
</tr>
<tr>
<td>Scott et al.</td>
<td>1995</td>
<td>23.3 (18.9, 28.2)</td>
<td>98.4 (95.9, 99.6)</td>
<td>14.5</td>
<td>0.78</td>
</tr>
<tr>
<td>Tanbo et al.</td>
<td>1992</td>
<td>46.3 (35.0, 57.8)</td>
<td>100 (71.5, 100)</td>
<td>462.5</td>
<td>0.54</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>25.9 (23.0, 29.0)</td>
<td>98.1 (96.1, 99.1)</td>
<td>13.8</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Note: LR+ = likelihood ratio of a positive test = sensitivity/(1 − specificity); if specificity was 100%, the value of 99.9% was used to calculate LR+; LR− = likelihood ratio of a negative test = (1 − sensitivity)/specificity.

criterion. Because all the studies in this review required a test with high specificity (such that an abnormal test reliably would indicate failure to conceive with treatment), a high FSH cutoff was chosen. This choice was based on either an a priori ROC analysis in the researchers’ laboratory, or a cutoff value that had been previously published. The cutoff values widely differed between some studies, however, due to the use of different FSH assays along with varying standards (42). Assay techniques ranged from RIA, to fluoroimmunoassay, and to enzyme immunoassay. Preparation standards varied between using the Second International Reference Preparation (2nd IRP) for human menopausal gonadotropin (hMG) and the World Health Organization (WHO) 2nd IRP (78 out of 549) for human FSH. Our ROC analysis however, confirmed that these cutoffs did not correlate with test properties. In addition, despite the different cutoffs, no significant heterogeneity was discovered in the overall results.

The sensitivities of both basal FSH and the CCCT are low, but can likely be improved without sacrificing their specificities. Adding a basal E₂ measurement will improve the sensitivity because, in some instances, a “normal” basal FSH will be falsely negative because the E₂ is actually elevated and actively suppressing the FSH level (20). Sensitivity can be further increased by repeating a “normal” test in a subsequent menstrual cycle because basal FSH and E₂ often fluctuate from normal to abnormal in consecutive menstrual cycles (43).

The “disease” prevalence (i.e., “no pregnancy with treatment” rate) that is commonly seen in infertility population ranges from 40%–100%, which corresponds to a “pregnancy with treatment” rate of 60%–0%. Due to this high “disease” prevalence in the screening population, both basal FSH and the CCCT had high PPVs and low NPVs in this “disease” prevalence range. Changes in the PPV are highly dependent on changes in either test specificity or “disease” prevalence (i.e., changes in sensitivity have a minimal effect on the PPV) (41). This explains why the PPVs for both tests were similar, even though their sensitivities were different. Changes in the NPV are highly dependent on changes in either the test sensitivity or “disease” prevalence. The NPVs for both tests in the “disease” prevalence range of 40%–100% were similarly low due to a combination of poor sensitivity and high “disease” prevalence.

Some clinicians feel that it is not valuable to screen infertility patients who are less than 35 years old (because these women would be at low risk for the “disease”) (44), whereas others have disagreed (45, 46). If we assume that these potentially “good prognosis” patients have a 60% chance of “pregnancy with treatment” (or a 40% chance of “no pregnancy with treatment”), and adjust the analysis to this hypothetical prevalence, both basal FSH and the CCCT would still have an excellent PPV (91.7% and 90.1%, respectively).

Systematic reviews are subject to limitations resulting from selection bias, publication bias, and the methodological quality of the original studies (47). In our analysis, selection bias was unlikely because we conducted a broad and exhaustive search, and only needed to exclude four studies that did not present adequate data for abstraction purposes. We did include two studies on basal FSH that only reported results with the outcome of live births (24, 26). This could introduce a bias because there may have been some pregnancies in patients with an abnormal basal FSH that ultimately did not result in a live birth. The total number of patients in these two studies (394), however, was a small fraction of the overall group (6,296), and the two studies together accounted for only 11% of the weight in the overall estimate. Excluding these studies did not cause a material change in our findings.

Although publication bias cannot be eliminated, all the published studies were consistent. Most but not all studies adopted similar screening criteria. It is still possible that there could be other unaccounted differences (e.g., race or ethnicity) among the studied populations that could bias the outcomes. In addition, more aggressive or varying infertility treatments may have been employed in subjects from different studies. These potential sources of bias, however, would apply to both the basal FSH and CCCT groups.

In conclusion, basal FSH and the CCCT are similar in predicting the ability to achieve a clinical pregnancy in women undergoing infertility treatment. With either test, a normal result is not useful, but an abnormal result virtually confirms that pregnancy will not occur with treatment. Basal FSH should be preferred over the CCCT because it is sim-

![Figure 2](Image 46x715 to 287x739)
pler, more cost-effective, and free from the potential for the adverse effects of CC.

References
10. Richardson SJ, Senikas V, Nelson JF. Follicular depletion during the
11. Sackett DL, Haynes RB, Tugwell P. A basic science for clinical