Diagnostic Efficiency of Home Pregnancy Test Kits

A Meta-analysis

Lori A. Bastian, MD, MPH; Kavita Nanda, MD, MHS; Vic Hasselblad, PhD; David L. Simel, MD, MHS

Objective: To assess the diagnostic efficiency of home pregnancy test (HPT) kits.

Data Sources: A literature search of English-language studies was performed with MEDLINE and a review of bibliographies.

Study Selection: Studies were included if HPT kits were compared with a criterion standard (laboratory testing), if they used appropriate controls, and if data were available to determine sensitivity and specificity.

Data Extraction: Two investigators independently extracted data, and disagreement was resolved by consensus. Sensitivity, specificity, and an effectiveness score (a measure of the discriminatory power of the test, with higher scores implying greater effectiveness) were calculated.

Data Synthesis: Five studies evaluating 16 HPT kits met the inclusion criteria. The range of sensitivities for HPT kits was 0.52 to 1.0. In studies where urine samples obtained by the investigators were tested by volunteers, sensitivity was 0.91 (95% confidence interval [CI], 0.84-0.96). However, the sensitivity was less in studies where subjects were actual patients who performed the test on their own urine samples (sensitivity, 0.75 [95% CI, 0.64-0.85]). The test effectiveness score was 2.75 (95% CI, 2.3-3.2) for studies where subjects were volunteers but deteriorated to 0.82 (95% CI, 0.4-1.2) for studies with actual patients.

Conclusions: The diagnostic efficiency of HPT kits is greatly affected by characteristics of the users. Despite the popularity of these kits, the relatively low effectiveness scores of these kits when used by actual patients are of concern. We suggest that manufacturers of HPT kits publish results of trials in actual patients before marketing them to the general public.

Arch Fam Med. 1998;7:465-469

HOME PREGNANCY test (HPT) kits have become increasingly popular since the first kit was released in the mid-1970s. These kits currently make up the fastest-growing segment of the home-diagnostic testing market. In the United States, approximately 33% of women have used an HPT kit to determine their pregnancy status before seeking professional health care. Most studies have found that women choose to use HPT kits because of the speed of obtaining results and the convenience of testing at home. Another advantage of the HPT kit is that the woman is the first person to know that she is pregnant. Since some women prefer to wait until they are sure they are pregnant before visiting their physician, HPT kits may lead to an earlier pregnancy diagnosis. An earlier diagnosis provides an opportunity for health care providers to counsel women about pregnancy options and to discourage potentially harmful behaviors, such as smoking and use of alcohol or drugs. The history of HPT kits parallels the development of laboratory tests for urinary human chorionic gonadotropin (HCG). The first kits used chemical and hemagglutination-inhibition methods, but most current kits use HCG-directed monoclonal antibodies. The active ingredients in monoclonal-based kits are the HCG α-chain–specific monoclonal antibodies, the β-chain–specific antibody/enzyme conjugate, the chromogenic substrate solution, and buffer solution. In the presence of urine HCG, the monoclonal antibody binds the hormone and produces a reaction, usually a color change because of the chromogenic substrate and buffer solutions. A reaction should not occur when HCG is absent, because the antibody adheres only to HCG. The accuracy of HPT kits is claimed to be 97% to 99% by the manufacturers. The newer products, such as Advance (Advanced Care Products, Ortho Pharmaceutical Corp, Raritan, NJ), Answer (Carter Products, Carter-Wallace, Inc, New York, NY), Clearblue (Unipath Diagnostics Co, New York, NY), and many others, have been advertised heavily in the mass media and on television. However, the diagnostic efficiency of these kits is unknown.
MATERIALS AND METHODS

DATA SOURCES

We searched the MEDLINE and HEALTHSTAR databases for English-language articles concerning the diagnosis of pregnancy that were published between 1966 and 1996. The key words used were pregnancy, diagnosis, pregnancy tests, and home tests. The HPT kits were not developed before 1966. References cited in articles and those listed in the bibliographies of standard obstetric texts were also retrieved. Articles were systematically reviewed by 2 of us (L.A.B. and K.N.) and given a grade of A through C based on the study design and level of evidence.17 Studies were included if the results of the HPT kit under investigation were compared to a criterion standard (laboratory tests) and used appropriate controls, and if data were available to determine sensitivity and specificity with total sample size greater than 20 (attempts were made to reach authors of potential articles to obtain additional information needed to determine sensitivity and specificity). We also attempted to obtain information from manufacturers of HPT kits but were unsuccessful.

STUDY SELECTION

Through the MEDLINE, textbook reference, and bibliography searches, we initially identified 53 articles; 43 were excluded either because the article was a review or because the HPT kit was not compared with a criterion standard laboratory-based urine or serum HCG test. The remaining 10 articles (Table 1) were then analyzed and 5 more were excluded.15,16,18-25 The additional exclusions were because the study had no control group of nonpregnant patients,3 there were insufficient data for determining sensitivity and specificity,18,19 the kit was no longer available because of its demonstrated poor performance,20 or the study had an inadequate sample size.21

DATA EXTRACTION

Data were abstracted independently by 2 of us (L.A.B. and K.N.) by means of structured forms that were pretested. Disagreements were resolved by consensus. Sensitivity, specificity, and a test effectiveness score were calculated. The test effectiveness score has been used in previous meta-analyses because it allows comparison of the relative and absolute ability of tests to discriminate those with from those without the target condition.26 A logistic odds transformation of the sensitivity and specificity allows creation of more normally distributed frequency plots of the test results for pregnant vs nonpregnant women. The effectiveness score quantifies the degree of overlap between the 2 plots and is interpreted directly as the number of SDs separating the means of the 2 curves. Tests that lead to considerable overlap between the 2 plots would have effectiveness scores of 1.0 or less and would not effectively distinguish pregnant from nonpregnant women.27 An effectiveness score of 1.0 means that 27% of pregnant women have tests results equivalent to those of women who are not pregnant. Tests that lead to minimal overlap between the 2 plots would be highly efficient in distinguishing pregnant from nonpregnant women and would have effectiveness scores approaching 3.0 and greater. Thus, a pregnancy test with an effectiveness score of 3.0 would yield results for pregnant women that are 3 SDs away from those of a population of nonpregnant women; the overlap in frequency plots for tests with an effectiveness score of 3.0 is only 3% of the patient sample. A test of homogeneity of both sensitivity and test effectiveness score was performed to evaluate consistency of findings across studies in these 2 categories. Because studies were found to be heterogeneous (P < .05), data were analyzed statistically by empirical Bayesian methods to arrive at summary statistics of sensitivity and test effectiveness score with 95% confidence intervals (CIs).28

RESULTS

General characteristics of the 10 studies retrieved initially are presented in Table 1. The sensitivity, specificity, and test effectiveness scores of the 5 retained studies (16 kits) are presented categorized by the HPT kit in Table 2. These studies achieved methodologic quality scores of either A or B. Three of the studies evaluated volunteers who performed the pregnancy tests on study samples obtained previously by the investigators. Two of the studies evaluated pregnancy tests performed by women who collected their own urine samples according to the kit instructions and performed the pregnancy test on their own samples.

The summary sensitivity was 0.91 (95% CI, 0.84-0.96) for studies where subjects were volunteers. Test performance deteriorated in studies where subjects were women who collected and tested their own samples, as demonstrated by a decreased summary sensitivity of 0.75 (95% CI, 0.64-0.85). Effectiveness scores of HPT kits also differed between these 2 groups. The Figure shows the...
effectiveness scores with 95% CIs, stratified by whether the study used volunteers or patients as subjects. The pooled test effectiveness score approached the desired benchmark value of 3.0 for studies in which volunteers performed the test (pooled effectiveness score, 2.75 [95% CI, 2.3-3.2]). However, kits were inefficient when women collected their own urine and performed the tests themselves (pooled test effectiveness score, 0.82 [95% CI, 0.4-1.2]).

COMMENT

These findings demonstrate differences in performance of the HPT kits. The low sensitivity and effectiveness score when HPT kits were used by women evaluating their own samples suggests that consumers and physicians should be concerned about the diagnostic efficiency of these kits, especially when the test result is negative for pregnancy. Despite the potential for problems with the HPT kits, most women and their physicians consider them reliable. Clinicians routinely advise patients to use these kits before scheduling prenatal appointments.3 Some physicians also rely on the results of HPT kits before treating patients with potentially teratogenic medications.

The overall marketing success of HPT kits led to the development of other home testing kits, such as an ovulation test kit and a human immunodeficiency virus test kit.29 On the basis of our review, we suggest that manufacturers of all home testing kits for any target condition should publish results of trials in actual patients before marketing them to the general public. If there are differences between volunteers and actual users, then modifications should be required until performance meets acceptable standards.

With one third of pregnant women using HPT kits, the low sensitivity (high rate of false-negative results) is a public health concern. False-negative results, even if they occur 10% of the time, may result in a delay in obtaining proper prenatal care and a missed opportunity to potentially motivate change in behaviors such as smoking or use of alcohol or drugs.13 A false-negative result may affect the feasibility and safety of pregnancy termination.

Two major reasons exist for the high false-negative rate when testing is performed by women on their own samples. First, women may be obtaining their samples...
In a 1993 study from France, 27 HPT kits were studied for their diagnostic efficiency. The investigators found a sensitivity range for all kits in France of 3% to 100%. They then tested the best 11 kits in 638 inexperienced volunteers. Of the 478 positive (pregnant) urine samples distributed, 230 were falsely interpreted as negative (sensitivity, 48%). The main explanation for the high rate of false-negative results was difficulty in understanding the instructions in the HPT kits, regardless of the socioeconomic situation (age, education, and employment) of the subjects. Valanis and Perlman, who studied only pregnant women, found that only 32% of users complied with all test kit instructions. The incidence of false-negative results in this study was 24.3%.

The fewer false-positive results with current monoclonal-based kits have been attributed to ectopic sources of HCG or elevated levels of circulating luteinizing hormone. Ectopic HCG production may rarely occur with certain tumors, such as small-cell carcinoma of the lung. High levels of luteinizing hormone may occur in postmenopausal women and in women just before ovulation. Proteinuria does not interfere with monoclonal-based kits, but it can result in inconclusive readings in the hemagglutination-inhibition tests. A few medications, such as methadone hydrochloride, carbamazepine, and aspirin, as well as medical conditions, such as ovarian cysts, abscesses, and pelvic inflammatory disease, may also interfere with the hemagglutination-inhibition test results. False-positive results are not thought to be as significant a health concern as false-negative results, as they should lead to a prenatal appointment and follow-up laboratory testing. False-positive results, however, can have extremely devastating psychological effects on the woman and her significant other.

A population of particular concern with regard to using HPT kits is teenagers. A recent study of teenagers requesting pregnancy tests in health departments showed that 28% of adolescents had used an HPT kit before their visit. Of those teenagers who were pregnant, one third had at least 1 negative pregnancy test before their posi-

### Table 2. Diagnostic Characteristics of Kits by Study

<table>
<thead>
<tr>
<th>Source, y</th>
<th>Type of Kit*</th>
<th>Sensitivity, %</th>
<th>Specificity, %</th>
<th>Effectiveness Test Score (95% CI)†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baker et al,22 1976</td>
<td>Ova II</td>
<td>52.6</td>
<td>51.9</td>
<td>0.09 (−0.5 to 0.7)</td>
</tr>
<tr>
<td>Arends and Udall,23 1977</td>
<td>Predictor</td>
<td>97</td>
<td>96</td>
<td>3.4 (2.6 to 4.2)</td>
</tr>
<tr>
<td>Doshi,16 1986</td>
<td>Answer</td>
<td>78</td>
<td>64</td>
<td>1.2 (0.4 to 1.6)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>Answer 2</td>
<td>82</td>
<td>64</td>
<td>1.0 (0.3 to 1.9)</td>
</tr>
<tr>
<td>Hicks and Iosefsohn,24 1989</td>
<td>e.p.t.</td>
<td>82</td>
<td>75</td>
<td>1.1 (0.5 to 1.7)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>e.p.t. plus</td>
<td>90</td>
<td>92</td>
<td>2.5 (1.9 to 3.1)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>Advance</td>
<td>86</td>
<td>91</td>
<td>2.2 (1.6 to 2.8)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>Fact</td>
<td>100</td>
<td>93.5</td>
<td>3.7 (2.1 to 5.3)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>e.p.t.</td>
<td>100</td>
<td>100</td>
<td>3.6 (2.1 to 5.2)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>Answer 2</td>
<td>100</td>
<td>93.5</td>
<td>3.6 (2.0 to 5.2)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>First Response</td>
<td>92.9</td>
<td>100</td>
<td>3.3 (1.8 to 4.9)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>Advance</td>
<td>91.2</td>
<td>100</td>
<td>3.3 (1.8 to 4.8)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>e.p.t.</td>
<td>88.2</td>
<td>100</td>
<td>3.2 (1.7 to 4.7)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>Predictor</td>
<td>100</td>
<td>76.7</td>
<td>2.9 (1.4 to 4.4)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>Daisy 2</td>
<td>97.5</td>
<td>60.5</td>
<td>2.0 (1.1 to 3.0)</td>
</tr>
<tr>
<td>Latman and Bruot,25 1989</td>
<td>Acu-test</td>
<td>51.7</td>
<td>88.9</td>
<td>1.3 (0.5 to 2.1)</td>
</tr>
</tbody>
</table>

*See Table 1 for kit manufacturer names and locations.
†CI indicates confidence interval.

Effectiveness scores for the home pregnancy test kits according to whether the kits were tested by patients or volunteers. The dashed line indicates an effectiveness score of 3.0 would yield results for pregnant women that are 3 SDs away from those of a population of nonpregnant women.

Before the recommended number of days after their first missed menstrual period (usually 9 days), when HCG levels become reliably detectable by the kits. Although many kits advertises its effectiveness at 9 days after the user’s last menstrual period, the sensitivities reported by manufacturers (≥90%) are not applicable until 2 weeks after the last menstrual period.

Another reason for false-negative results is operator error. Operator errors result from failure to read or follow instructions, or difficult procedures inherent to the kit. In 1 study, pharmacy students evaluated the instruction leaflets in 9 popular kits. Although they did not rate the instruction leaflets significantly differently across kits, they did determine that the results of the 3 kits that used a color change were easier to interpret than the other kits. In a 1993 study from France, 27 HPT kits were studied for their diagnostic efficiency. The investigators found a sensitivity range for all kits in France of 3% to 100%.

©1998 American Medical Association. All rights reserved.
tive result. The decision by a sexually active teenager to test herself for pregnancy marks the need for counseling about contraception, even when the result is negative. This has led others to recommend discouraging teenagers from using HPT kits so that those with negative results performed in clinics will be afforded the opportunity to discuss health behaviors intended to reduce the rate of teenage pregnancy. It is unlikely that we can prevent teenagers from using HPT kits. An alternative suggestion would be to encourage manufacturers to label kits with a warning suggesting that teenagers talk with an adult about their pregnancy test result, even if it is negative.

Publication bias may impair our ability to assess home testing kits because manufacturers of self-diagnostic kits do not publish their results. An editorial by the vice president of an HPT manufacturer cited that extensive, but unpublished, clinical trials in hundreds of women were conducted before the kit was marketed. We attempted to obtain these unpublished data from manufacturers of HPT kits but were unsuccessful. We do not know if publication bias would change our findings, but usually publication bias results in the increased publication rates of studies with good results. We also are unaware of whether the issue of actual patient use vs testing by volunteers is adequately evaluated in premarking trials.

CONCLUSIONS

Researchers have been concerned about the differences in diagnostic test characteristics when HPT kits were used outside of a controlled laboratory setting. In the hands of experienced technicians, the HPT kits have been proven to be almost as accurate (97.4%) as professional laboratory testing. However, they are less accurate when performed by consumers. The limitation of self-testing is the ability of the users to perform the test. It is essential that HPT kits provide adequate instructions that are easy to read and understand. Our study suggests that HPT kit instructions should be reviewed to (1) make sure women understand them; (2) encourage women to wait at least 2 weeks after a missed period before performing the test; and (3) notify women of the potential for false-negative results.

Clinicians should be concerned about the diagnostic efficiency of HPT kits, given the relatively low effectiveness score when used by actual patients. When a patient calls reporting a negative result, she should be encouraged to repeat the test 1 week later if she remains amenorrheic, and to call her provider if the test result remains negative. When a patient calls reporting a positive result, she should be encouraged to schedule an appointment for her first prenatal visit to confirm that she is pregnant. Because most manufacturers do not publish the results of their trials in actual patients, we are not able to report sensitivity, specificity, and test effectiveness scores for all HPT kits currently on the market. Consumer Reports ranked these kits on the basis of the manufacturer’s reported accuracy, ease of use, and cost and determined Answer to be the best value. Without more information from the manufacturers, we cannot recommend 1 specific HPT kit. Further research is needed in this area.

REFERENCES


Accepted for publication October 6, 1997.
Presented in part at the annual sessions of the Society of General Internal Medicine, Washington, DC, May 1, 1997.
Reprints: Lori A. Bastian, MD, MPH, Durham Veterans Affairs Medical Center, 508 Fulton St (152), Durham, NC 27705 (e-mail: LBastian@acpub.duke.edu).