CASES IN PRIMARY CARE LABORATORY MEDICINE

Testing pitfalls and summary of guidance on sex hormone testing

W Stuart A Smellie

consultant, Clinical Laboratory,
General Hospital, Bishop Auckland
DL14 6AD
info@smellie.com

BMJ 2007;334:91-4
doi: 10.1136/bmj.39038.614317.AE

Sex hormone testing in women is difficult to interpret, and to produce valid conclusions tests must be timed correctly and follow a rational sequential testing strategy

Laboratories see large differences in the use of tests to investigate subfertility. The use and interpretation of tests for the female sex hormones can be problematic; although population reference data exist, values change markedly within and between the different phases of the menstrual cycle, and values outside of quoted reference ranges often provide very limited information and can be misleading. This is particularly true for women with cycles that are irregular or atypical in length. Good examples of this can be seen in both menopause and subfertility.

This paper presents two typical clinical cases involving biochemical investigation of female reproductive capacity and menopause and considers the further investigations that may or may not be helpful in these situations. It also presents a summary of the relevant researched guidance.

Case 1

A 47 year old woman with no gynaecological history apart from two uncomplicated term pregnancies in her 20s presented to her general practitioner with a three month history of tiredness, 2.5 kg weight gain, and low mood with occasional flushing. On questioning, she reported that she had had irregular and occasionally missed periods over the previous five months, and she had no specific features of an affective disorder. She had not had a period in the previous six weeks.

Thyroid function tests and gonadotrophin concentrations were requested and reported as thyroid stimulating hormone 6.8 mIU/l (reference range 0.4-5.5 mIU/l), free thyroxine 14 pmol/l (reference range 10-23 pmol/l), luteinising hormone 7.2 IU/l, follicle stimulating hormone 3.7 IU/l, and oestradiol 168 nmol/l. The table shows cycle phase reference ranges.

A full blood profile and renal profile were unremarkable. The patient was advised that she had borderline low thyroid function but that her symptoms indicated that she could be entering the menopause, although her tests did not indicate this. She was asked to return in

SUMMARY POINTS

- Hormone testing has very limited use in diagnosing the menopause
- Monitoring patients on hormone replacement therapy relies on clinical response in most situations
- Early investigation of subfertility is recommended only in specific groups of patients
- A detailed menstrual history is essential to interpret gonadotrophin results
- A luteal progesterone concentration above 20-30 nmol/l seven days before onset of menses effectively excludes endocrine related subfertility; intermediate concentrations may necessitate a repeat test

two months for a further thyroid test, or sooner if her symptoms worsened.

Four weeks later the patient presented again after developing more troublesome flushing and increased tiredness. Repeat thyroid function tests were superimposable on the initial set; thyroid antibodies were negative, follicle stimulating hormone was 17.3 IU/l, and luteinising hormone was 14.2 IU/l. The patient was told that she was probably entering the meno-

Table 1 | Reference data for female sex hormones*

Hormone	Reference range
Follicle stimulating hormone (IU/l)	
Follicular phase	3.5-12.5
Ovulatory	4.7-21.5
Luteal phase	1.7-7.7
Postmenopausal	25.8-134.8
Luteinising hormone (IU/l)	
Follicular phase	2.4-12.6
Ovulatory	14.0-95.6
Luteal phase	1.0-11.4
Postmenopausal	7.7-58.5
Oestradiol (pmol/l)	
Follicular phase	46.0-607
Ovulatory	315-1828
Luteal phase	161-774
Postmenopausal	<18.4-201

*Adult female, Roche ECLIA method for Elecsys E170 module, Roche Diagnostics, Indianapolis, IN, USA.

This is the eighth article in this series

pause and was started on combined oral hormone replacement therapy.

When she was reviewed eight weeks later, she complained of persistent flushing and tiredness and her gonadotrophins and oestradiol were measured to assess response to the hormone replacement therapy. The patient had started a menstrual period six days previously.

The hormone concentrations were reported as follicle stimulating hormone 7.7 IU/l, luteinising hormone 5.2 IU/l, and oestradiol 2425 pmol/l. The doctor telephoned the laboratory to ask what action was needed for the high oestradiol. This conversation revealed that the woman had been started on a preparation containing conjugated equine estrogens, and no further measurements were recommended. The patient's symptoms improved with a modest rise in the dose of hormone replacement therapy.

Case 2

A 28 year old woman presented to her general practitioner complaining of difficulty in becoming pregnant. She had no gynaecological history apart from a termination of pregnancy five years before and had been taking the contraceptive pill since then, until deciding to try to become pregnant 10 months previously. She did not complain of any menstrual irregularities and had no other specific signs or symptoms apart from being overweight (body mass index 29). As she was concerned, her doctor requested progesterone and gonadotrophin measurement (on day 20 of her cycle) and a semen sample from her partner. The progesterone concentration was reported as 16.7 nmol/l, "unlikely to be consistent with ovulation"; luteinising hormone was 27.2 IU/l, and follicle stimulating hormone was 15.3 IU/l. Her partner's semen sample had a normal sperm count and motility.

Further investigations two weeks later revealed a serum testosterone of 1.5 nmol/l (female reference range 0.7-3 nmol/l), a prolactin of 186 mIU/l (non-pregnant reference range 60-619 mIU/l), and normal thyroid function. A pelvic ultrasound scan showed no evidence of polycystic ovarian syndrome.

When she returned to see her doctor after a further two weeks, the patient reported that she had missed her period, which had been due two weeks previously, and had done a pregnancy test, which was positive. On further questioning, she admitted that her periods, although regular, tended to be long at around 35 days after she stopped the oral contraceptive pill. The pregnancy progressed successfully to term.

Discussion

These cases illustrate some of the difficulties in measuring female sex hormones. These difficulties apply particularly in women with menstrual irregularities, as results that may be entirely physiological in one phase of the cycle may be very abnormal for another.

Case 1

The presenting symptoms in a 47 year old woman were consistent with but not diagnostic of the onset of meno-

pause. Gonadotrophin testing is rarely helpful in this situation, as normal results do not exclude menopause and raised concentrations, although suggestive, do not necessarily confirm it.¹ Raised concentrations also do not indicate an inability to conceive, which is based on clinical and not laboratory criteria (contraception is needed for 12 months of amenorrhoea in patients over 50 years old and for 24 months in those under 50¹). Fluctuating gonadotrophin results are common in the perimenopause, and consensus guidance is that gonadotrophin testing may only be of help in atypical presentations or in women under 45, and serial measurements over time may be needed to add supporting information in making a diagnosis. It may be a relevant test when amenorrhoea cannot be used as a marker, as in women taking the contraceptive pill or those who have had a hysterectomy, although as the contraceptive pill suppresses gonadotrophin results a low concentration may be impossible to interpret.

Similarly, very little guidance exists to support oestradiol or gonadotrophin testing in women receiving hormone replacement therapy, except possibly oestradiol to assess compliance or transcutaneous absorption from a patch, or before implant replacement in women with an estrogen implant to avoid excessive accumulated concentrations.² Oestradiol concentrations can seem alarmingly high in women receiving conjugated equine estrogens because of assay cross reaction and do not provide useful monitoring information in women on oral hormone replacement therapy.

In uncomplicated classic presentations, both the diagnosis and monitoring of menopausal symptoms is clinical, and laboratory measurements may often serve only to introduce confusion. Where laboratory testing is considered necessary in unusual presentations, follicle stimulating hormone alone is sufficient; if it is raised consistently over 20-40 IU/l (depending on method and author^{1,3}) in the absence of periods, it is usually indicative of ovarian failure but is no guarantee of sterility.

Case 2

Investigation of subfertility is not advised in the absence of predisposing factors (box 1) or in women under 35 until failure to conceive after 12 months^{4,5}; reassurance is generally sufficient. In this case, the patient's concern prompted earlier testing.

Although this woman volunteered no menstrual complaints, the later finding that she had long menstrual cycles meant that the day 20 measurement of serum

Box 1 | Who should be investigated for subfertility testing?

We recommend investigation of both partners of couples who have been unable to achieve pregnancy:

- After one year of regular unprotected intercourse
- Earlier where a history of predisposing factors (such as amenorrhoea, oligomenorrhoea, pelvic inflammatory disease, or undescended testes) exists
- Earlier where a woman is aged 35 years or over
- Earlier if attempting pregnancy is associated with other risks (such as changing treatment for epilepsy)

progesterone was unreliable, as she had not progressed sufficiently far into the luteal phase for her progesterone to reach a concentration conventionally reported as being consistent with ovulation.

Conventionally, a day 21 progesterone is recommended to confirm ovulation. More accurately, consensus guidance recommends a progesterone measurement seven days before the expected onset of menses.⁵ More than one measurement may be needed if cycles are of irregular length. A progesterone concentration below 20 nmol/l is taken to indicate that ovulation is unlikely, and a concentration over 30 nmol/l suggests that ovulation is very likely to have occurred. By default, therefore, a grey area seems to exist, which may necessitate repeat testing.

Further investigation by measurement of gonadotrophins in samples early in the cycle, typically on days two to four,^{6,7} is only warranted if a progesterone sample indicates a poor luteal phase, unlikely to be consistent with ovulation. Conventional guidance includes oestrogen measurement, although the added information this offers is limited. Additional thyroid function and prolactin measurements are recommended in this situation only if thyroid or pituitary disease is suspected.^{4,5} The first gonadotrophin sample that was actually taken on day 20 of a 35 day cycle in this patient returned a high luteinising hormone result compared with the follicle stimulating hormone, initially suggesting polycystic ovarian syndrome.^{8,9} This actually reflects the mid-cycle peri-ovulation gonadotrophin surge that occurs later in a long cycle. Further investigation could have been avoided with a day 28 progesterone sample, which would have been expected to be consistent with ovulation, and it seems that the patient in fact became pregnant in this cycle, highlighting the importance of

Box 2 | Recommended initial tests for investigation of subfertility in men and women

What tests should be used to investigate subfertility in men?

We recommend

- Clinical examination, followed by
- Seminal fluid analysis
- Serum luteinising hormone, follicle stimulating hormone, and testosterone if repeated azoospermia is present

What tests should be used to investigate subfertility in women?

- Test of ovulation:
- Mid-luteal phase serum progesterone (seven days before the expected onset of menses, depending on cycle length: day 21 in a 28 day cycle)
- Further investigation of ovulatory failure when indicated by serum progesterone or in women with amenorrhoea/oligomenorrhoea without suspected androgen excess:
- Follicle stimulating hormone, luteinising hormone, oestradiol
- Thyroid function tests and prolactin if suggestive signs or symptoms present
- Additional test if androgen excess suspected (for example, polycystic ovarian syndrome):
- Testosterone (or free androgen index)

Box 3 | Investigation of secondary amenorrhoea

When should secondary amenorrhoea be investigated?

We recommend investigation

- After cessation of menstruation for six consecutive months in a woman who had previously had regular cycles
- Unless age and symptoms are compatible with menopause
- And when pregnancy has been excluded

What tests should be used to investigate secondary amenorrhoea?

We recommend

- Pregnancy test
- Follicle stimulating hormone, luteinising hormone
- Prolactin and thyroid function tests if suggestive signs or symptoms are present
- Testosterone/free androgen index in hirsute women or women with other features of androgen excess

reassurance and the need for further time to allow a patient to conceive.

Questions and answers: learning points

The series of questions and answers summarised in the boxes can be found in the second and fourth reviews of best practice in primary care pathology published in the *Journal of Clinical Pathology*.^{1,3} The boxes present the key recommendations from these reviews.

What are the sources of evidence?

As can be seen from some of the ranges and thresholds quoted, published gonadotrophin concentrations representing "thresholds" for ovulatory failure and progesterone concentrations consistent with ovulation vary considerably, although good agreement exists about the situations in which testing is or is not considered useful. Some of these differences may reflect differences in methods at the time the work was done, but they may also reflect the difficulty in establishing absolute thresholds for a dynamic endocrine response, in which absolute values can reflect a range of factors from precise timing of the sample to the individual patient's endocrine axis. Gonadotrophin and progesterone concentrations can vary by a factor of 10-30 or more during the menstrual cycle, suggesting that any estimate of a threshold would have a large confidence interval. Nevertheless, correct timing of specimens and a rational approach to testing will allow the great majority of patients with underlying endocrine disease responsible for subfertility to be identified, avoiding the need for further endocrine investigation of women who are likely to be ovulating. Similarly, the limited utility of gonadotrophin testing to identify menopause and its inability to predict sterility in perimenopausal patients is consistent with the guidance that it should generally not be used in this situation.

I thank Susan Richardson for typing the manuscript; the clinical practice section of the Association of Clinical Biochemists (in particular D B Freedman and I S Young); R Gama (Association of Clinical Pathologists) and R Neal and N Campbell (Royal College of General Practitioners), who kindly reviewed the original work; and R Gosling, E Kalu, E Sherriff, and C Croucher, who were authors of the original guidance.

USEFUL WEBSITES

- See first article in this series (*BMJ* 2006;333:83-6, doi: 10.1136/bmj.333.7558.83) for a list of general websites
- British Fertility Society (www.britishfertilitysociety.org.uk/humanfertility/index.html)—information source for health and science professionals involved in human fertility

Competing interests: None declared.

- 1 Smellie WSA, Forth J, Sundar S, Kalu E, McNulty CAM, Sherriff E, et al. Best practice in primary care pathology: review 4. *J Clin Pathol* 2006;59:893-902.
- 2 Gangar K, Cust M, Whitehead MI. Symptoms of oestrogen deficiency associated with supraphysiological plasma oestradiol concentrations in women with oestradiol implants. *BMJ* 1989;299:601-2.
- 3 Smellie WSA, Forth J, McNulty CAM, Hirschowitz L, Lilic D, Gosling R,

et al. Best practice in primary care pathology: review 2. *J Clin Pathol* 2006;59:113-20.

- 4 National Institute for Clinical Excellence. *Fertility: assessment and treatment for people with fertility problems*. London: NICE, 2004. (Clinical guideline 11.)
- 5 Royal College of Obstetricians and Gynaecologists. *The initial investigation and management of the infertile couple*. London: RCOG Press, 1998.
- 6 Scott R, Toner J, Muasher S, Oehninger S, Robinson S, Rosenwaks Z. Follicle-stimulating hormone levels on cycle day 3 are predictive of in vitro fertilization outcome. *Fert Steril* 1989;51:651-4.
- 7 Kim YK, Wasser SK, Fugimoto VJ, Klein NA, Moore DE, Soules MR. Utility of FSH:LH ratio in predicting reproductive age in normal women. *Hum Reprod* 1997;12:1152-5.
- 8 Fox R, Conigan E, Thomas PA, Hull MG. The diagnosis of polycystic ovaries in women with oligomenorrhoea: predictive power of endocrine tests. *Clin Endocrinol* 1991;34:127-31.
- 9 Robinson S, Rodin DA, Deacon A, Wheeler MJ, Clayton RN. Which hormone tests for the diagnosis of polycystic ovarian syndrome. *Br J Obstet Gynaecol* 1992;99:232-8.

Accepted: 9 November 2006

CLINICAL EPIDEMIOLOGY NOTES

What is heterogeneity and is it important?

John Fletcher

RESEARCH p82

clinical epidemiologist
BMJ, London WC1H 9JR
jfletcher@bmj.com

BMJ 2007;334:94-6

doi: 10.1136/bmj.39057.406644.68

Three simple examples from recent *BMJ* papers illustrate the importance of heterogeneity in a systematic review and how readers can assess it

Three systematic reviews published in the *BMJ*, including one in this issue, have referred to heterogeneity and dealt with it in three different ways.^{1 2 3} So what is heterogeneity, and how do we assess its importance in a systematic review?

Clinical heterogeneity

Sometimes trials are just looking at different concepts. Reviewers might set out to summarise interventions for improving patients' ability to make treatment choices; the trials, however, might have covered diverse interventions, such as information leaflets, CD Roms, counselling sessions with a nurse, and training in consultation techniques for doctors. Although the interventions try to achieve the same end result (to improve patients' ability to make choices), they are different in nature.

In theory, we could add all the trials in this review together and come up with a number, but would this be useful? Would the averaged number apply to all these diverse interventions? The interventions are so different that combining them does not make clinical sense. This is an example of clinical heterogeneity. Other circumstances that may give rise to clinical heterogeneity include differences in selection of patients, severity of disease, and management. Judgments about clinical heterogeneity are qualitative, do not involve any calculations, and can be made by putting forward a convincing argument about similarities (or differences) between the trials.

Statistical heterogeneity

Individual trials in a systematic review may seem to measure the same outcome but may have results that are not consistent with each other. Some trials show a benefit while others show harm, or the trials are inconsistent in the size of benefit or harm. This is the case in the systematic review of medications to prevent allergic reactions caused by contrast media.¹ The trials that measured effects on cutaneous symptoms of allergy showed a range of odds ratios from 0.12 favouring the medication to 1.02 favouring the control (fig 1). This is an example of statistical heterogeneity.

How can you detect it and does it matter?

Statistical heterogeneity is apparent only after the analysis of the results. Heterogeneity may be judged graphically (by looking at the forest plot) and be measured statistically. In a forest plot from the systematic review of calcium supplementation,² the error bars for each trial include the summary result, which suggests that statistical heterogeneity is not a problem and that the message is a consistent one (fig 2).

To determine whether significant heterogeneity exists, look for the P value for the χ^2 test of heterogeneity. A high P value is good news because it suggests that the heterogeneity is insignificant and that one can go ahead and summarise the results. Because statistical tests for heterogeneity are not very powerful it is sensible to use a higher P value than usual (say, $P > 0.1$) as the cut-off for a decision and to think about clinical heterogeneity anyway.

The systematic review of calcium supplementation passes the test, and the authors have rightly summarised the effects on bone density using a simple fixed effects model. This model assumes that all trials

This is the first in a series of occasional articles explaining statistical and epidemiological tests used in research papers in the *BMJ*.