Clinical Study

Receiver operating characteristic analysis of the performance of basal serum hormone profiles for the diagnosis of polycystic ovary syndrome in epidemiological studies

Héctor F Escobar-Morreale, Miryam Asunció, Rosa M Calvo, José Sancho and José L San Millán
Departments of Endocrinology and Molecular Genetics, Hospital Ramón y Cajal, Carretera de Colmenar km. 9'100, 28034 Madrid, Spain
(Correspondence should be addressed to H F Escobar-Morreale; Email: hector.escobar@uam.es)

Abstract

Objective: We have used receiver operating characteristic (ROC) analysis to determine the diagnostic performance of several serum parameters, in order to evaluate their potential usefulness in establishing the diagnosis of polycystic ovary syndrome (PCOS) in epidemiological studies.

Design: Prospective study.

Methods: One hundred and fourteen women reporting spontaneously for blood donation were included in the study. Menopausal and oral contraceptive-treated women were excluded. Serum samples were obtained at the moment of donation, independently of fasting, time of day or day of menstrual cycle. Measurements included total testosterone, sex hormone-binding globulin (SHBG), dehydroepiandrosterone sulfate (DHEAS), LH, FSH and estradiol. The free testosterone (FT) concentration and the free androgen index (FAI) were calculated from testosterone and SHBG levels. ROC curves were calculated for all these serum determinations.

Results: Eight patients were diagnosed with PCOS, according to the presence of oligomenorrhea, hirsutism, acne and/or hyperandrogenemia, and exclusion of non-classic congenital adrenal hyperplasia, hypothyroidism and hyperprolactinemia. Of the parameters studied SHBG, FAI, FT and DHEAS were considered adequate measures for the diagnosis of PCOS. For example, serum SHBG levels showed an area under the ROC curve of 0.875 (S.E. 0.045 (95% confidence interval 0.800–0.929). A SHBG decision threshold, 37 nmol/l had a sensitivity of 87.5%, a specificity of 86.8%, a positive likelihood ratio of 6.63, and a negative likelihood ratio of 0.14, for the diagnosis of PCOS.

Conclusions: Our present results strongly suggest that decreased SHBG levels, and increased FAI, free testosterone concentration and DHEAS concentrations, are highly effective as single analytical procedures in epidemiological studies for the detection of PCOS in women of reproductive age.

European Journal of Endocrinology 145 619–624

Introduction

Polycystic ovary syndrome (PCOS) is one of the more prevalent endocrine disorders in women, affecting 4–7% of women of reproductive age (1–3).

The widely accepted research criteria for the diagnosis of PCOS are derived from the conference sponsored in 1990 by the National Institute of Child Health and Human Development (NICHD) (4). Most of the participants in this conference agreed that PCOS diagnosis requires, in order of importance: (i) clinical and/or biochemical hyperandrogenism, (ii) oligo-ovulation, and (iii) exclusion of other etiologies such as non-classic congenital adrenal hyperplasia, hyperprolactinemia or androgen-secreting tumors (4). Detection of polycystic ovaries by ultrasound was not considered essential for the diagnosis of PCOS (4), and later studies have confirmed the limited value of sonographic parameters for the prediction of the hormonal abnormalities characteristic of PCOS (5).

The broad application of these criteria has permitted significant advances in our knowledge of PCOS, such as the precise evaluation of its prevalence (1, 2), or the recognition of PCOS as a familial syndrome (6, 7).

However, because no indications were given in the NICHD conference on how to define the individual criteria (4), this ambiguity may be responsible in part for the frequent discrepancies found in the literature regarding the prevalence and pathophysiology of PCOS.

In a clinical setting, the use of NICHD criteria is easy. Ovulation can be monitored using basal body temperature charts or serum progesterone measurement on days 22–24 of the menstrual cycle, and sampling for serum androgens can be controlled for the day of the menstrual cycle or the hour of the day.
On the contrary, in population-based studies involving large number of subjects, the NICHD criteria for the diagnosis of PCOS are difficult to apply, a problem that is especially evident in epidemiological and familial studies. The accuracy of the NICHD criteria may decrease if several observers are involved in the evaluation of hirsutism and clinical hyperandrogenism, and the methods needed for the precise evaluation of ovulatory function are not usually applicable for large transversal studies.

In this context, appropriate use of non-subjective variables, such as some serum variables likely to be related to PCOS, may facilitate and improve the diagnosis of this condition. However, the performance of serum hormone determinations for PCOS detection in that setting has not been studied, and most familial and epidemiological studies rely only on serum testosterone levels for diagnosing biochemical hyperandrogenism and PCOS (3, 8, 9).

Receiver operating characteristic (ROC) curves are used in the evaluation of diagnostic test performance (10). ROC curves are a graphical representation of the relationship between the sensitivity and the specificity of a diagnostic test over all possible diagnostic cut-off values (10). In addition to the evaluation of test performance, ROC curves are used to choose the cut-off values for diagnostic tests, either to obtain maximum sensitivity, maximum specificity, or a balance of both.

ROC analysis has been used to evaluate the diagnostic performance of serum hormone determinations for PCOS detection in a clinical setting, namely, the discrimination of PCOS patients from other hyperandrogenic patients and from healthy controls, under controlled sampling conditions (11, 12).

On the contrary, the diagnostic performance of serum determinations for PCOS detection in epidemiological studies has not been established. For this reason, in the present study we have performed ROC analysis, considering NICHD criteria as the ‘gold-standard’ for the diagnosis of PCOS, of the basal serum concentrations of several hormones and related parameters usually measured in these patients, in order to evaluate their potential usefulness as single diagnostic tests for screening for this disorder in epidemiological studies.

Materials and methods

Subjects, protocol and assays

One hundred and fifty-four consecutive women were recruited prospectively, after giving informed consent, to participate in a study to evaluate the prevalence of PCOS, and hirsutism, in Spain (1). These women were blood donors reporting spontaneously to the Blood Bank of the Hospital Ramón y Cajal. For the present study, women in whom no serum samples were obtained, and those women taking oral contraceptives, were excluded, the latter because of the profound impact of oral contraceptive treatment in steroidogenic and gonadotropin hormone profiles.

One hundred and fourteen women were finally included in the present study. The presence of hirsutism, acne and menstrual dysfunction were recorded as described in detail elsewhere (1). Serum samples were obtained at the moment of blood donation, and therefore these samples were not controlled for fasting, time of day or day of menstrual cycle.

Serum was assayed for total testosterone, sex hormone-binding globulin (SHBG), dehydroepiandrosterone sulfate (DHEAS), luteinizing hormone (LH), follicle-stimulating hormone (FSH) and estradiol (E2). The free testosterone (FT) concentration and the free androgen index (FAI) were calculated from testosterone and SHBG concentrations, as previously described (1, 13). In women presenting with clinical and/or biochemical evidence of hyperandrogenism and menstrual dysfunction, serum prolactin and 17-hydroxyprogesterone measurements excluded hyperprolactinemia and non-classic congenital adrenal hyperplasia (1). Serum concentrations of the above variables were assayed as described elsewhere (1).

Women were diagnosed with PCOS when presenting with menstrual dysfunction, clinical and/or biochemical hyperandrogenism, and exclusion of other etiologies. Menstrual dysfunction was considered when the women had oligomenorrhea, defined by more than six cycles per year with a length of more than 36 days (14), and/or when the patient had not had any menstrual bleeding for three consecutive months, during the last year. Clinical hyperandrogenism was defined by the presence of hirsutism, represented by a hirsutism score of 8 or more (15), by persistence of acne during the third decade of life or later, or by the presence of androgenic alopecia. Hyperandrogenemia was defined by a circulating testosterone, FT, DHEAS and/or FAI higher than the 95th percentile of the women studied here who had no clinical evidence of hyperandrogenism, or menstrual disturbances, and who were not taking hormonal medication nor had had previous oophorectomy or hysterectomy. Normal serum concentrations of prolactin (<24 μg/l), thyrotropin (0.5–4.5 mIU/l) and 17-hydroxyprogesterone (<6.05 nmol/l) excluded hyperprolactinemia, thyroid dysfunction and non-classic 21-hydroxylase deficiency in all the women who achieved the other criteria for the diagnosis of PCOS.

Informed consent was obtained from every patient and control. The study was approved by the Institutional Review Board of the Hospital Ramón y Cajal, and was conducted according to principles expressed in the Declaration of Helsinki.

ROC analysis

For ROC analysis, patients presenting with PCOS were considered affected, and the remaining women were considered non-affected.
ROC curves are constructed by plotting the sensitivity (true-positive fraction) on the ordinate as a function of the complement of specificity (false-positive fraction), for all the possible cut-off values of the diagnostic test (10). Therefore, the more deviated towards the left upper corner the curve is, the higher the sensitivity and the specificity the diagnostic test has for all possible cut-off values or, namely, the higher the adequacy of the diagnostic test for disease detection (16).

As opposed to accuracy, sensitivity (or 1−(false-negative fraction)) and specificity (or 1−(false-positive fraction)) are not dependent on the prevalence of the disease in the sample. Thus, ROC curve analysis provides a description of disease detectability, which is independent from both disease prevalence and decision threshold effects (16).

The area under the ROC curve (AURC), which can be calculated using the trapezoidal rule, characterizes ROC curves. The AURC represents the probability of correctly distinguishing between affected and non-affected subjects (17). Therefore, the perfect diagnostic test, not having false-positive or false-negative results, would have an AURC of 1 and, on the contrary, tests with AURCs ≤0.5 would not discriminate between affected and non-affected individuals.

The Wilcoxon statistic, W, is used to test the null hypothesis that the diagnostic test cannot be used to discriminate between affected and non-affected subjects (i.e. the probability equals 0.5), and is also used to predict the statistical properties of the AURC (17). For ROC analysis, the most important property is the standard error (s.E.w), allowing the calculation of the 95% confidence interval (CI) (AURC±1.96 s.E.w) (17).

ROC curve analysis, and the statistical significance of the differences in the AURC of the ROC curves of the different serum determinations, were calculated using ROCKIT 0.9β software (Dr Charles E Metz, Department of Radiology, University of Chicago, Chicago, IL, USA). The unpaired t-test or the Mann–Whitney U–Wilcoxon W test was used to compare PCOS patients and non-affected subjects, as appropriate. P < 0.05 was considered statistically significant.

Results

Eight of the 114 women included in the studied were diagnosed with PCOS. For ROC curve analysis, the eight patients with PCOS were considered as affected, and the remaining 106 women were considered as non-affected. In the latter group, there were two women presenting with hirsutism, and another 12 women had acne, but their menstrual cycles were regular, and therefore they were not considered to have PCOS.

Table 1 Clinical and hormonal variables in PCOS subjects and non-affected subjects. PCOS was defined according to NICHD endocrine criteria (4). Result are means±S.D. Serum samples were not controlled for fasting, time of day or day of menstrual cycle.

<table>
<thead>
<tr>
<th></th>
<th>PCOS (n = 8)</th>
<th>Non-affected subjects (n = 106)</th>
<th>P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>25.0±5.4</td>
<td>34.5±9.4</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>25.5±4.2</td>
<td>24.2±3.1</td>
<td>0.280</td>
</tr>
<tr>
<td>Waist-to-hip ratio</td>
<td>0.80±0.07</td>
<td>0.80±0.06</td>
<td>0.952</td>
</tr>
<tr>
<td>Age at menarche (years)</td>
<td>13.1±1.5</td>
<td>12.5±1.5</td>
<td>0.396</td>
</tr>
<tr>
<td>Number of pregnancies</td>
<td>0.8±0.9</td>
<td>1.4±1.4</td>
<td>0.223</td>
</tr>
<tr>
<td>Number of abortions</td>
<td>0.1±0.4</td>
<td>0.3±0.8</td>
<td>0.541</td>
</tr>
<tr>
<td>Hirsutism score</td>
<td>6.8±7.8</td>
<td>2.3±2.3</td>
<td>0.096</td>
</tr>
<tr>
<td>SHBG (nmol/l)</td>
<td>32.3±17.7</td>
<td>68.4±38.6</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>FAI</td>
<td>6.5±5.1</td>
<td>2.4±2.4</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>DHEAS (µmol/l)</td>
<td>11.6±4.2</td>
<td>6.7±3.8</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>Calculated FT (pmol/l)</td>
<td>33.8±22.0</td>
<td>16.2±14.0</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>Total testosterone (nmol/l)</td>
<td>1.7±0.9</td>
<td>1.3±0.8</td>
<td>0.065</td>
</tr>
<tr>
<td>E2 (pmol/l)</td>
<td>217±148</td>
<td>331±256</td>
<td>0.204</td>
</tr>
<tr>
<td>LH (IU/l)</td>
<td>8.1±5.2</td>
<td>10.2±11.0</td>
<td>0.803</td>
</tr>
<tr>
<td>FSH (IU/l)</td>
<td>7.4±2.9</td>
<td>11.8±18.2</td>
<td>0.816</td>
</tr>
</tbody>
</table>

*The unpaired t-test or the Mann–Whitney U–Wilcoxon W test was used to compare PCOS patients and non-affected subjects, as appropriate.
differences between these AURCs, with the exception of a higher AURCs for the FAI as compared with FT \( z = -3.599, P < 0.001 \). However, our study did not have statistical power to detect such small differences (i.e. at least 557 subjects are needed to detect differences between AURCs of 0.875 and 0.825 with a 90\% power).

Serum testosterone, \( E_2 \), LH and FSH levels had small AURCs, which were statistically lower as compared with those of SHBG, FAI, FT and DHEAS (data not shown), possibly reflecting the variability of serum sampling with respect to the day of the menstrual cycle.

The ROC curves of SHBG, FAI, FT and DHEAS are represented in Fig. 1, and the suggested decision thresholds for the detection of PCOS using these measurements are summarized in Table 3. The best combination of sensitivity (0.875) and specificity (0.868) was that of SHBG, using a decision threshold of <37 nmol/l. In fact, only one of the eight patients with PCOS had SHBG values \( \geq 37 \) nmol/l whereas 14 of the 106 non-affected women had SHBG levels <37 nmol/l.

**Discussion**

For years, research into PCOS has been burdened by the different criteria used for its definition. The adoption of the endocrine criteria introduced after the NICHD conference in 1990 (4) has permitted significant advances in our knowledge of this disorder in the last decade. Among other milestones, the broad adoption of these criteria has allowed the recognition of PCOS as an extremely frequent disorder, that occurs in 4–7\% of women of reproductive age (1, 2), presenting with familial association (8), and associated with significant reproductive, metabolic and cardiovascular morbidity (18).

However, the NICHD criteria for the diagnosis of PCOS may be difficult to apply, especially in situations involving large numbers of women, such as in familial and epidemiological studies. On the one hand, there is no consensus on the particular definition of these criteria. Clinical and/or biochemical evidence for hyperandrogenism may be interpreted differently; some authors require hyperandrogenemia for PCOS diagnosis (3, 9), whereas the presence of hirsutism or acne is enough for others (1, 2). Also, the precise diagnosis of oligo-ovulation is difficult, requiring techniques (basal body temperature charts, luteal phase serum progesterone levels) that are not usually available for epidemiological studies. For these studies, most authors use oligomenorrhea as indicative of oligo-ovulation (1, 2, 9), but again there is no consensus regarding the length of the menstrual cycle or the number of menstrual cycles per year required for the definition of oligomenorrhea.

**Table 3** Decision thresholds for the detection of PCOS as defined by NICHD endocrine criteria (4).

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Decision threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive likelihood ratio</th>
<th>Negative likelihood ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHBG</td>
<td>&lt;37 nmol/l</td>
<td>0.875</td>
<td>0.868</td>
<td>6.63</td>
<td>0.14</td>
</tr>
<tr>
<td>Calculated FT</td>
<td>( \geq 23 ) pmol/l</td>
<td>0.625</td>
<td>0.849</td>
<td>4.14</td>
<td>0.44</td>
</tr>
<tr>
<td>FAI</td>
<td>( &gt;3.67 )</td>
<td>0.750</td>
<td>0.858</td>
<td>5.30</td>
<td>0.29</td>
</tr>
<tr>
<td>DHEAS</td>
<td>( &gt;11.6 ) ( \mu )mol/l</td>
<td>0.625</td>
<td>0.915</td>
<td>7.36</td>
<td>0.41</td>
</tr>
</tbody>
</table>

www.eje.org
Our present results suggest that serum determinations might be enough for the screening of PCOS in epidemiological studies. This approach has the advantage of not relying on qualitative or semi-quantitative parameters for PCOS diagnosis, such as hirsutism scores, presence or absence of acne, or number and length of menstrual cycles, which are difficult to standardize. Moreover, this approach may avoid clinical evaluation of the subjects, which usually needs special training of the research team.

Accordingly, we have evaluated the adequacy of several serum parameters to discriminate between women with or without PCOS in an unselected population-based sample using ROC analysis, applying NICHD criteria as the ‘gold-standard’ for PCOS definition.

Not surprisingly, ROC curve analysis suggests that F AI, FT and DHEAS measurements are adequate for this purpose, as hyperandrogenemia is one of the criteria for the diagnosis of PCOS. However, serum SHBG concentration appears to be equally useful, and especially adequate to detect women with PCOS. The AURC for SHBG was 0.875 (0.800–0.929), namely, serum SHBG has a 0.875 probability of correctly discriminating between women with PCOS and those without this syndrome. This figure is comparable to those of widely accepted diagnostic tests, such as serum thyrotropin IRMA assays for the diagnosis of hyperthyroidism (19).

We chose a serum SHBG level <37 nmol/l as decision threshold for the diagnosis of PCOS, because of its high sensitivity and specificity, and its elevated positive likelihood ratio. Although the AURC for SHBG was not statistically higher than the AURC of the other serum hormone levels, the balance of sensitivity and specificity at the suggested SHBG <37 nmol/l decision threshold was the best of all the parameters tested, suggesting that SHBG might be the ideal parameter for PCOS detection in epidemiological studies.

The performance of serum SHBG for the diagnosis of PCOS is in agreement with its pivotal role in the pathophysiology of this syndrome. SHBG synthesis and secretion are down-regulated by androgens and, specifically, by increased portal insulin levels (20). Insulin resistance, resulting in compensatory hyperinsulinism and reduced SHBG levels, is central to the development of PCOS (21).

In our series, reduced SHBG levels were specifically associated with PCOS, and did not result from the known association of PCOS with obesity (22), because the body mass index and the waist-to-hip ratio (and index of central adiposity, and of insulin resistance (23)) were virtually identical in PCOS patients and non-affected individuals.

Furthermore, serum SHBG, FT and DHEAS concentrations and the F AI allowed PCOS detection in our study independently of sampling conditions, including fasting (critical for evaluation of insulin sensitivity), time of day (which may affect serum adrenal steroid levels), or day of the menstrual cycle (critical for serum gonadotropins and E2), and this is a major advantage in situations in which sampling conditions are difficult to control, such as in epidemiological or familial research studies.

Conversely, the absence of control for sampling conditions in our present study probably explains the poor performance of serum gonadotropins and E2, certainly affected by the wide variations of these hormones during the menstrual cycle, and is possibly related to the relative inadequacy of testosterone for PCOS detection. On the contrary, those androgens that do not change markedly during the menstrual cycle, such as DHEAS, or the indexes of FT calculated from SHBG, were also useful for PCOS detection.

ROC analysis has several advantages for the evaluation of the adequacy of diagnostic tests, including that it is not dependent on the prevalence of the disorder to be detected, and does not depend on the decision threshold employed (16). However, ROC analysis has limitations, mainly related to the definition of the independent criteria by which the diagnosis is made, bias selection of subjects with complete sets of the diagnostic tests to be compared, and appropriateness of the sample to the clinical scenario in which the diagnostic test is going to be used (10).

In our study the independent criteria for PCOS definition were those of the NICHD conference of 1990, and all the subjects had the same serum parameters measured, thus avoiding selection bias. However, our present results should only be applied to epidemiological or population-based studies, in which only a small fraction of the population is expected to have PCOS.

Because of these considerations, it is important to emphasize that our present results cannot be extrapolated to clinical practice, in which serum hormone determinations are used to differentiate PCOS from other forms of hyperandrogenism, and sampling conditions are usually controlled.

For this purpose, two recent studies have used logistic regression analysis and ROC analysis to determine the performance of several serum measurements in detecting PCOS in a clinical setting. Turhan et al. (11) evaluated prospectively the performance of serum LH, FSH, testosterone, FT, SHBG and insulin concentration in a group of 32 consecutive women with PCOS, defined by NICHD criteria and ultrasonography, and 25 healthy controls. The best fit was obtained by including LH/FSH ratio, insulin and testosterone concentrations in the model, but only LH/FSH ratio and insulin retained significance in the forced entry method in logistic regression analysis (11). Serum SHBG was not included in the model (11). The authors then performed ROC analysis of LH/FSH ratio and insulin levels, obtaining AURCs of 0.823±0.032 and 0.813±0.037 respectively (11), values that are comparable to those we have found in our present study. Only by combining LH/FSH ratio and insulin levels did they obtain an AURC of 0.929±0.063 (11).
In the second study, Koskinen et al. (12) showed that an estimate of the probability of PCOS, derived from the levels of the variables retained in the logistic regression model (serum LH, FSH and androstenedione concentrations), had the highest clinical utility for discriminating PCOS patients and controls in a clinical setting. LH, LH/FSH ratio, testosterone and androstenedione serum determinations had an AURC above 0.900 (12). Surprisingly, serum SHBG levels had a low AURC, and were not useful for PCOS detection in their series (12). However, this study had significant differences from our present study and from that of Turhan et al. (11); women with PCOS were recruited retrospectively – a potential source of bias – and the diagnosis of PCOS was based on the presence of oligomenorrhea and sonographic polycystic ovaries, instead of relying on NICHD criteria (12).

In conclusion, performance of serum SHBG, FT, FAI and DHEAS measurements for the detection of PCOS in epidemiological studies is very high even when sampling conditions are not controlled. If our preliminary results are confirmed by future studies in larger population-based studies, the screening of PCOS may be based on the determination of serum parameters, instead of relying in inaccurate clinical criteria.

Acknowledgements

The authors thank Dr Elena Lasa, Dr Concepción Zamora and all the nurses from the Blood Bank of the Hospital Ramón y Cajal, Madrid, Spain, for their collaboration in recruitment and sampling of the population studied. Ms Mª Paz Muñoz kindly helped with hormonal measurements. The authors also thank DPC-DIPESA, Madrid, Spain, for supplying free of charge the reagents used for measurements of serum testosterone, SHBG, LH, FSH and E2. This study was supported by a grant (Proyecto 08.6/0022/1998 to H F E-M) from the Consejería de Investigación and by a grant (Proyecto 08.6/0022/1998 to R M C) from the Fondo de Investigación (Proyecto FIS 00/0414 to H F E-M, and Contrato FIS 624 to R M C). This study was supported by a grant (Proyecto 08.6/0022/1998 to R M C) from the Fondo de Investigación. The authors thank Dr Elena Lasa, Dr Concepción Zamora and all the nurses from the Blood Bank of the Hospital Ramón y Cajal, Madrid, Spain, for their collaboration in recruitment and sampling of the population studied. Ms Mª Paz Muñoz kindly helped with hormonal measurements.

References