EXTERNAL QUALITY ASSESSMENT OF ASSAYS
FOR HORMONES AND TUMOR MARKERS IN ITALIAN LABORATORIES

G.C. ZUCCHELLI (a), A. PILO (a), R. MALVANO (a), C. SIGNORINI (b), G.F. BOLELLI (c)
and A. ALBERTINI (b)

(a) Istituto di Fisiologia Clinica, CNR, Pisa
(b) Istituto di Chimica, Facoltà di Medicina, Università degli Studi, Brescia
(c) Fisioterapia della Reproduzione, CNR, Bologna

Summary. - External quality assessment (EQA) programs run by CNR/Tecnostandard for immunoassays of hormones and tumor markers, started in 1980, presently include as many as 20 analyses; about 300 laboratories are involved in these programs. For all immunoassays submitted to the EQA, the inspection of cumulative results allows the current situation to be documented for total variability and its within-kit and between-kit components (the former accounting for the reproducibility and robustness of the kits and the latter for their systematic differences of estimation). For 13 assays subjected to EQA for longer, the variability trends over time are depicted, and single factors affecting the overall quality of particular assays are identified. Among these, experimental simplification of kit structure, alignment of calibrators with an acknowledged reference material, and adoption of monoclonal-antibody based two-site assays can be mentioned. On the contrary, neither automation of the procedures nor (more expectedly) increasing use of nonisotopic techniques has proved effective in significantly improving the analytical quality.

KEY WORDS: external quality assessment, assay of hormones, assay of tumor markers.

Riassunto (Valutazione esterna di qualità dei dosaggi di ormoni e marcatori tumorali in laboratori italiani). - I programmi di valutazione esterna di qualità (VEQ), condotti da CNR e Tecnostandard per i dosaggi di ormoni e marcatori tumorali, iniziati nel 1980, includono oggi 20 analisi; sono circa 300 i laboratori a cui oggi complessivamente coinvolti. Per ognuno degli analisi presi in esame, l'osservazione dei risultati cumulativi consente di documentare la situazione attuale e quanto concerne la variabilità totale e le sue componenti "entro-kit" e "fra-kit" (la prima riguarda la reproducibilità e "robustezza" dei kit, la seconda le loro differenze sistematiche di stima). Per i 13 dosaggi con una più lunga storia di VEQ, vengono illustrati gli andamenti della variabilità nel tempo e, per alcuni casi specifici, vengono identificati fattori singoli con effetti critici sulla qualità complessiva di prestazione: fra questi, la semplificazione della struttura operativa dei kit, l'allineamento dei calibratori a materiale di riferimento riconosciuto, la diffusione di metodi a due siti basati su anticorpi monoclonali. Al contrario, né l'automazione del procedimento analitico né (più prevedibilmente) il crescente impiego di tecniche non isotopiche hanno dimostrato un qualche effetto significativo nel migliorare la qualità del dosaggio.

PAROLE CHIAVE: valutazione esterna di qualità, dosaggio di ormoni, dosaggio di marcatori tumorali.

Introduction

External quality assessment (EQA) schemes largely adopted in chemical industry have been extended, over the last decade, to immunoassays. Giving participant laboratories a tool for controlling their analytical performance is regarded as the primary goal of EQA. A further function, however, that EQA permits is of great interest. EQA results can provide, retrospectively, a realistic information on the quality of assays, as actually done under routine conditions. This information proved useful both for laboratories to improve their performance and for kit manufacturers to stimulate setting up of more reliable products.

Beginning in 1980, a national EQA, sponsored by CNR/Tecnostandard (*), has been run for immunoassays of hormones [1-3] and then of tumor markers [4]; in this paper, the state of the art and the trends of the assay results submitted in EQA are tentatively portrayed through in-

(*) Pilot EQA programs are conducted by the Italian National Research Council (CNR), while routine programs are run as a service by Tecnostandard, a company operating in the field of immunoassay quality assurance. In both cases, the same scientific committee is charged with the management of the EQA schemes and data treatment.

Fig. 2 - Assays of steroid hormones: trends of total variability (CV, %) throughout the EQA periods. EQA periods: (1) 1981/2, (2) 1982, (3) 1983, (4) 1984/1, (5) 1984/2, (6) 1985, (7) 1986, (8) 1987/1, (9) 1987/2, (10) 1988/1, (11) 1988/2, (12) 1989/1, (13) 1989/2, (14) 1990/1. For definition of the total variability, see [7].


Fig. 4 - Correlation between the total variability (CV, %) and frequency of use of “direct” procedures in steroid immunoassays. Three assays are considered as a group. The data shown refer to the period 1981-1987.
EQA outline

The CNR/Tecnostandard EQA for assays of endocrinological and oncological interest includes at present as many as 20 analytes (14 hormones, 6 tumor markers); cumulatively, about 300 laboratories are involved. The scheme does not substantially differ from similar programs (see e.g. [5, 6] in this issue); participants, supplied monthly with 3 to 6 unknown samples (freeze-dried or stabilized sera, mailed at room temperature), are asked to perform the assay routinely and to return results indicating the kit used. The data collected are computer-processed and periodic and end-of-period reports are prepared and distributed to participants. The periodic report contains a statistical evaluation of the results for each single EQA sample (mean, median, standard deviation, coefficient of variation, range) and a histogram of the results of all the laboratories identified by a code. The end-of-period report (or cumulative report) includes an estimate of the average bias (inaccuracy) and average imprecision achieved by each laboratory in assaying all the samples dispatched in

the EQA cycle (usually a six-month period during which 12-18 samples are assayed); in addition, the end-of-period report contains estimates of the analytical performance of the kits more frequently used in the survey, as obtained from the cumulative data relative to the whole cycle.

The approaches followed for the analysis of EQA data are discussed in detail elsewhere in this issue [7].

Immunoassay performance: the present situation

An updated picture of immunoassay quality emerges from the EQA results obtained in the first six-month period of 1990. The variabilities observed are shown in Table 1; for each analyte included in the EQA the following statistics are reported:

- the total variability (related to concentration ranges where a roughly uniform spread of data might be assumed), which reflects both laboratory imprecision and between-kit, between-laboratory systematic discrepancies;

- the within-kit and between-kit components of the total variability derived by analysis of variance (ANOVA) (see details in [7]); the within-kit component reflects the average imprecision of the kits (including the batch-to-batch variability) and accounts for difficulties in assay management, whereas the between-kit component is related to the existence of systematic differences mainly arising from different antibody specificities and/or incorrect standard calibration.

For the majority of analytes concerned, the total variability ranges from 17 to 25% with some exceptions in both directions. Lower spreads are apparent for T4, cortisol, and E3 assays (seemingly due to the relatively high analyte concentration in the sample) and a larger one for LH, thus confirming a general situation already referred to in this issue [5].
Fig. 7 - Between-laboratory imprecision profile of TSH assay, found in different EQA periods in which the kits involved were exclusively based on polyclonal competitive technique (period 1982-1985, closed circles) or on MoAb “high sensitivity” two-site technique (period 1988-1989, open circles), respectively.

Table 1. - Assay variability from EQA data (1989-1990) (a)

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Concentration range</th>
<th>Total variability CV (%)</th>
<th>Components of variability (b) CV&lt;sub&gt;within&lt;/sub&gt; (%)</th>
<th>CV&lt;sub&gt;between&lt;/sub&gt; (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T3</td>
<td>1.4 - 4.5 nmol/l</td>
<td>16.9</td>
<td>12.1 (51)</td>
<td>11.8 (49)</td>
</tr>
<tr>
<td>T4</td>
<td>77 - 278 nmol/l</td>
<td>11.2</td>
<td>10.4 (86)</td>
<td>4.3 (14)</td>
</tr>
<tr>
<td>TSH</td>
<td>1.1 - 11 mIU/l</td>
<td>21.7</td>
<td>17.0 (61)</td>
<td>13.6 (39)</td>
</tr>
<tr>
<td>LH</td>
<td>7.7 - 44 IU/l</td>
<td>34.8</td>
<td>21.2 (37)</td>
<td>27.6 (63)</td>
</tr>
<tr>
<td>FSH</td>
<td>7.5 - 26 IU/l</td>
<td>17.2</td>
<td>14.6 (72)</td>
<td>9.0 (28)</td>
</tr>
<tr>
<td>PRL</td>
<td>444 - 1108 mIU/l</td>
<td>20.1</td>
<td>17.8 (78)</td>
<td>9.4 (22)</td>
</tr>
<tr>
<td>E2</td>
<td>0.3 - 3.1 nmol/l</td>
<td>20.4</td>
<td>17.9 (77)</td>
<td>9.9 (23)</td>
</tr>
<tr>
<td>Progesterone</td>
<td>5.6 - 54 nmol/l</td>
<td>21.1</td>
<td>18.1 (74)</td>
<td>10.8 (26)</td>
</tr>
<tr>
<td>Testosterone</td>
<td>2.2 - 31 nmol/l</td>
<td>24.6</td>
<td>21.0 (73)</td>
<td>12.8 (27)</td>
</tr>
<tr>
<td>Cortisol</td>
<td>243 - 1103 nmol/l</td>
<td>14.6</td>
<td>13.0 (80)</td>
<td>6.6 (20)</td>
</tr>
<tr>
<td>Aldosterone</td>
<td>0.1 - 1.1 nmol/l</td>
<td>24.7</td>
<td>22.3 (82)</td>
<td>10.6 (18)</td>
</tr>
<tr>
<td>DHEA-S</td>
<td>3.0 - 22 µmol/l</td>
<td>20.9</td>
<td>19.9 (93)</td>
<td>6.3 (9)</td>
</tr>
<tr>
<td>E3</td>
<td>201 - 803 nmol/l</td>
<td>12.6</td>
<td>11.6 (86)</td>
<td>4.7 (14)</td>
</tr>
<tr>
<td>HPL</td>
<td>1.7 - 8.1 µg/ml</td>
<td>19.4</td>
<td>14.8 (58)</td>
<td>12.5 (42)</td>
</tr>
<tr>
<td>AFP</td>
<td>18.3 - 102 IU/l</td>
<td>17.3</td>
<td>15.2 (77)</td>
<td>8.1 (23)</td>
</tr>
<tr>
<td>CEA</td>
<td>8.0 - 38 µg/l</td>
<td>20.6</td>
<td>14.0 (46)</td>
<td>15.2 (54)</td>
</tr>
<tr>
<td>Ferritin</td>
<td>27 - 279 µg/l</td>
<td>18.9</td>
<td>14.8 (61)</td>
<td>11.7 (39)</td>
</tr>
<tr>
<td>CA19-9</td>
<td>15 - 95 U/l</td>
<td>20.5</td>
<td>18.3 (80)</td>
<td>9.2 (20)</td>
</tr>
<tr>
<td>CA125</td>
<td>24 - 178 U/l</td>
<td>19.1</td>
<td>16.5 (75)</td>
<td>9.7 (25)</td>
</tr>
<tr>
<td>CA15-3</td>
<td>16 - 120 U/l</td>
<td>18.8</td>
<td>17.3 (85)</td>
<td>7.3 (15)</td>
</tr>
</tbody>
</table>

(a) The data refer to 13-18 samples assayed during the most recent EQA cycles by 152-202 laboratories for TecnoStandard and 62-91 laboratories for CNR pilot schemes. 16 to 32 kits are involved for each analyte, except for the cases of CA19-9, CA125 and CA15-3 for which only 5 to 8 kits are used.

(b) Within-kit and between-kit components are indicated as CV<sub>within</sub> and CV<sub>between</sub>, respectively. The percent contribution of the within-kit component, i.e. (CV<sub>within</sub>/CV<sub>total</sub>)x100, and of the between-kit component (CV<sub>between</sub>/CV<sub>total</sub>)x100, are indicated in parentheses; the computational details of the breakdown of variability are reported elsewhere in this issue [7].
A historical overview

For the 13 assays with a longer EQA experience, the trends of total variability over time are depicted in Figs 1-3. Some peculiar behaviours are apparent at a glance, despite occasional fluctuations due to the prevalence in the various EQA periods of “difficult” samples (in respect to the concentration of the analyte in the EQA samples), and possibly to kit turnover in the laboratory. In particular, major evidence is provided for more or less pronounced improvements successively occurring in the performance of assays of E2, progesterone, testosterone, CEA, and TSH and, in contrast, for a marked increase in the variability of LH assays. It can also be seen that for T3 and T4 the overall quality remained substantially unchanged throughout the whole decade of EQA. The same holds for cortisol, at least from 1984 on. Less marked trends are shown for the other analytes.

The common tendency to improvement in E2, progesterone, and testosterone assays resulted in approximately halving of the associated total variability over the period 1981-1987. As confirmed by the correlation of Fig. 4, this trend may be explained by the progressive replacement of extraction procedures with the less demanding direct measurement in untreated sample [8] (currently used without exception by all participants in the EQA).

The large variability observed for CEA assay in 1985-1986 is due mainly to differences between the calibrators supplied with the kits, as far as the molecular structure of

---

**Fig. 8.** Breakdown of the total variability of LH and FSH assays into within-kit and between-kit components. The percent contribution to the total variability of each component is indicated under the appropriate bars.

The within-kit component was found to contribute most to the total variability in the majority of assays (the percent contribution ranging from 60 to 90%, in 17 cases out of 20). This suggests that the variability of data is due to poor robustness of the kits rather than to systematic differences of estimates entailed by the different methodological approaches. Such a consideration, quite obvious for the assays of the mucin antigens (CA19-9, CA125, CA15-3) where only few kits using the same reagents (antibodies and standards from Centocor, Malvern, USA) are concerned, is also valid when a large number of methodologically dissimilar kits is involved. Exceptions are represented in assays of T3 and CEA, for which the two variability components appear to be equally important, and particularly in LH assay, whose anomalous behaviour is further evidenced here by the much larger contribution of the between-kit differences.

---

**Fig. 9.** Comparison of LH values measured by a polyclonal radioimmunoassay and by four different MoAb based immunoradiometric kits. The values were computed from cumulative regression analysis of EQA data collected in the 1987-1989 cycles.
the protein preparation used as standard and/or the tier of these preparations are concerned [9, 10]. A striking example demonstrating this situation is shown in Fig. 5, where a dramatic reduction of the spread of estimates (from ca. 60 to 20 CV%) is evident after normalizing the participants’ results with respect to a common standard. The sudden improvement of the agreement of CEAs observed in 1987 is attributable to a better alignment of kit calibrators in consequence of the adoption of the international WHO standard IRP 73/401 as reference. This is confirmed by the ANOVA results relative to the two discrete EQA periods compared in Fig. 6. Rather than to the small decrease observed for the within-kit variability (from ca. 18 to 14 CV%), the lowering of the total variability is due to the marked reduction of the between-kit component (from ca. 30 to 15 CV%) which, however, remains the major variability factor, thus giving evidence for residual systematic differences in measuring serum CEA.

In the case of TSH assay, the total variability decreased continuously since the beginning of the program. A number of factors, difficult to substantiate, may give rise to this trend, but a decisive role was certainly played by the introduction and adoption of “high-sensitivity” two-site immunometric assays using MoAbs [6]. This is evidenced by the comparison of the improvement profiles resulting from the EQA data recorded before 1986 (when the kits used were exclusively based on polyclonal competitive techniques) and after 1988 (when almost all laboratories used MoAb “high sensitivity” techniques) (Fig. 7); it can be seen that the between-laboratory spread of results for EQA samples with TSH concentration below 5 mIU/l has been markedly reduced since noncompetitive assays have been introduced.

In contrast to the above situations, a sudden worsening of the total variability is apparent for LH assay from 1987 on, coinciding with the prevailing use of MoAb-based kits in EQA, which became available to Italian laboratories around 1987. The different reactivity to LH epitopes between polyclonal and monoclonal reagents (and between the individual MoAb employed by kit manufacturers) provides an obvious explanation, although FSH and PRL assays, which share the same situation, do not appear as vulnerable to this variability factor. The results of the ANOVA in two successive EQA periods are perfectly in keeping with the above explanation. As seen in Fig. 8, the analysis of the components of variability evidences that for LH assay (but not for FSH) the relative effectiveness of within-kit and between-kit contributions has reversed over time, and that the latter now largely predominates (as seen already in Table 1). As Fig. 9 shows, systematic discrepancies of results are found with different kits, and a tendency to produce lower estimates is demonstrated for MoAb-based assays, which can further disagree between one another. This gives rise to the complex distributions of EQA results, exemplified in Fig. 10, in comparison to the situations previously met, thus further complicating the treatment of data.
The situations described above clearly indicate that technological development in immunoassay plays a role in defining the overall analytical performance, as it certainly does in determining the rapid turnover of kits used by participants in the EQA [11]. In this context, the increasing use of nonisotopic assays and automated procedures should be mentioned. In particular, automation (most generally combined with nonradioactive tracers) might be thought to be potentially effective in reducing the spread of measurements through strict control of the experimental variables (volumes, times, temperature). In fact, in neither case are decided differences indicated by the imprecision/bias plots of Fig. 11, relating to thyroid hormones (nonisotopic assay accounting for about 28% of total results, automation for 27%), cortisol (nonisotopic assay 21%, automation 18%) and the tumor markers AFP and CEA (nonisotopic assay about 40%, automation about 27%). This situation is quite obvious for nonisotopic procedures which are not expected per se to affect the overall analytical quality. As for the automated methods, the data collected in the EQA to this time do not appear to be significantly more precise than those produced by the most reliable manual kits; these findings suggest a tendency of manufacturers of automatic systems simply to meet the practicability requirements of the laboratories, rather than to actually improve the quality of the assays, and achieve analytical goals.

Acknowledgements

This work was partially supported by CNR, Special Project “Biotecnologie e Biostimolazione”.

Review submitted on invitation by the Editorial Board of the Annali. Accepted for publication on: 1 February, 1991.

REFERENCES


