The U.S. National Reference Preparation for Alpha-Fetoprotein In Mid-Pregnancy Maternal Serum

Submitters: Charles B. Reimer, S. Jay Smith, and Thomas W. Wells, Immunological Products Branch, Center for Infectious Diseases, Centers for Disease Control, Public Health Service, U.S. Department of Health and Human Services, Atlanta, GA 30333

Reviewers: James E. Haddow, Acting Medical Director, Foundation for Blood Research, Scarborough, ME 04074
William M. Hunter, Medical Research Council, Immunoassay Team, Edinburgh EH2 QW, Scotland
Lawrence M. Killingsworth, Clinical Chemistry and Immunology Laboratories, Sacred Heart Medical Center, Spokane, WA 99220
Phillipe Sizaret, Division of Environmental Carcinogenesis, International Agency for Research on Cancer, Lyon Cedex 2, France

Assigned Editor: Judith A. Clayton-Hopkins, Abbott Laboratories, North Chicago, IL 60064 (formerly at Oncofetal Antigen Laboratory, Centers for Disease Control, Atlanta, GA 30333)

The World Health Organization (WHO)1 International Reference Preparation for Human Alpha-Fetoprotein (AFP; no. 72/225), (subsequently referred to as the WHO Standard), collaboratively evaluated and assigned a value of 100 000 int. units2 of AFP per (glass-sealed) ampoule, is a freeze-dried pool of human cord serum that is currently available on a limited basis as a calibrator for AFP assays (1–4). Recently, the U.S. National Committee for Clinical Laboratory Standards (NCCLS) Subcommittee on Alpha-Fetoprotein recommended (5) that the Centers for Disease Control (CDC) produce a candidate lyophilized calibrator for AFP in midpregnancy maternal serum that contained approximately 400 int. units of AFP per milliliter, in a normal human serum matrix. Additionally, the CDC should organize the collaborative evaluation and calibration of this candidate national reference preparation in international units, vis-à-vis the WHO Standard, and in mass units derived by consensus from estimates obtained from “local” calibrators in current use, including calibrators from representatives of all AFP-kit manufacturers who had premarket approval applications with the U.S. Food and Drug Administration (FDA) when this consensus evaluation was organized (June 1979).

These tasks have been completed and this report provides the basis for recommending this candidate standard as the U.S. National Reference Preparation for Alpha-Fetoprotein in Midpregnancy Maternal Serum (subsequently referred to as the U.S. Standard).

Materials

The U.S. Standard for AFP: Normal human sera stored below −35 °C, from more than 20 healthy donors for the Atlanta, GA, area, were obtained from the CDC Serum Bank, pretested, and found to be negative for hepatitis-B surface antigen by radioimmunoassay; negative for rheumatoid factor by Behring Diagnostics’ latex-agglutination tests; free from fibrinogen, fibrin, or fibrin breakdown products as judged by single radial immunodiffusion against an appropriate antiserum; and containing <10 int. units of AFP per milliliter, as determined by an immunofluorometric assay (6, 7). These normal human sera were pooled (total volume 5.4 L), and 43.5 mL of a pool of straw-colored normal placental cord serum was added to give a final AFP concentration of approximately 400 int. units/mL. This AFP-enriched normal human serum pool was then frozen and thawed twice, centrifuged (20 000 × g for 90 min), filtered through a 0.45-μm pore size membrane filter, refrozen, thawed, and sterile-filtered through a 0.45-μm membrane filter. Volumes of 0.50 mL were aseptically dispensed into approximately 10 000 borosilicate vaccine vials, lyophilized, and sealed with butyl rubber stoppers while still under reduced pressure. The mean weight and coefficient of variation (CV) of the dispensed serum in 24 randomly selected vials before lyophilization were 514.67 mg and 0.70%, respectively; the mean dry weight and CV of the contents of five randomly selected vials afterwards were 43.16 mg and 0.53%, respectively; the mean water content and SD as determined by the Karl Fischer residual moisture test of 85 randomly selected samples were 0.55% and 0.19%, respectively, of the total weight of lyophilized material.

The vial contents of the U.S. Standard were sterile, as determined by USP-approved methods of sterility testing, and free of hepatitis-B surface antigen, by radioimmunoassay.

Groups of 12 or more vials were stored in controlled-environment chambers at 4, 25, 35, 48, and 55 °C. Vials were

---

1 Nonstandard abbreviations: WHO, World Health Organization; AFP, alpha-fetoprotein; NCCLS, U.S. National Committee for Clinical Laboratory Standards; CDC, (U.S.) Centers for Disease Control; FDA, Food and Drug Administration.
2 The WHO-approved abbreviation for the international units assigned to WHO International Biological Standards and WHO International Biological Reference Preparations under the control of the WHO Expert Committee on Biological Standardization is IU, not the SI recommendation, int. unit. Additionally, most centers for maternal serum AFP screening in the U.S. and abroad customarily report their AFP results either in IU/mL or in ng/mL. Because a WHO International Biological Standard is used here as the primary standard, the Submitters, not wishing to violate well-established customs, strongly prefer use of the 'nonstandard' (from the viewpoint of IUPAC) abbreviations: IU/mL and ng/mL. The Editors of the AACC, however, in keeping with usage in other Proposed Selected Methods, have used throughout int. unit, which IUPAC and IFCC state "may be defined when the result of a measurement is not a part of a recognized kind of quantity having a definable dimension... Anyone may define an arbitrary unit; the more authoritative 'international unit' is defined by an international body" (Recommendation 1973), e.g., WHO.
transferred at various times from these chambers to \(-35^\circ C\) storage for later reconstitution and analysis in the same run, to assess the rate of thermal degradation of the preparation.

 AFP-free human male serum diluent: Sera from healthy men, obtained from the CDC Serum Bank and tested as above but selected to contain <3 int. units of AFP per millilitcr, were pooled and absorbed by adding purified horse anti-human AFP that was covalently bound to 1.1-\(\mu\)m-diameter polymethacrylate spheres (6). After removal of the solid-phase immunosorbent by centrifugation, the pool contained <1 int. unit of AFP per millilitcr. Enough of this diluent was provided to the 11 collaborators for them accurately to dilute the WHO Standard to 400 int. units/mL according to the following directions.

The WHO Standard for AFP: A sufficient number of ampoules of this international calibratcr for AFP (no. 72/225) were obtained from the WHO (through Evaluator P.S.) to provide three ampoules for each of the 11 collaborators in this study. Detailed instructions were given to (and followed by) each collaborator to reconstitute this lyophilized preparation by adding 2.00 mL of distilled water, then accurately adjust the volume per ampoule to 5.00 mL with the AFP-free human male serum diluent supplied by CDC, and then accurately dilute an aliquot of this 20 000 int. units/mL solution to 400 int. units/mL with the same diluent.

Methods

Protocol: The protocol for this study required that the two preparations be evaluated together against the collaborator’s “local” calibratcr (working curve) in mass units in each of three independent runs performed on separate days with freshly reconstituted specimens. Starting with the reconstituted candidate U.S. Standard and with the 400 int. units/mL solution of the WHO Standard (made up with the AFP-free serum supplied), collaborators made all subsequent dilutions of the two preparations independently, using their own diluents and series of dilutions. Collaborators reported all their results from each run in nanograms per millilitcr, as derived from their own mass calibrators, for six replicate assays for each of five dilutions they made of both preparations, usually covering at least a 16-fold total range of concentration that was independently selected by each collaborator to be approximately centered on the optimal measurement range (log scale) of each collaborators’ assay system (except collaborator H, who reported only a fourfold range for his five dilutions).

Eight of the 11 collaborators of this study used their own radioimmunoassay (RIA) reagents for AFP, two others (collaborators B and K) used an RIA kit supplied by the same manufacturer (not one of the above-mentioned eight), and the CDC laboratory (collaborator A) used its own reagents in an immunofluorometric assay for AFP (6, 7).

Statistical considerations: Outlier rules were designed to identify and remove: (a) within-collaborator, within-dilution mass concentration inconsistency—that is, situations in which a particular collaborator’s mass estimates at any particular dilution were highly variable as compared with the data of all the collaborators; and (b) within-collaborator, within-dilution, relative concentration inconsistency—that is, situations in which a particular collaborator’s estimates of the relative mass concentrations of equal dilutions of the two standards were highly variable in relation to data of all the collaborators.

For the first outlier rule, we first determined that the SDs of the set of six estimates (within-collaborator, -run, -preparation, -dilution) increased with their respective means such that the CVs tended to be constant. Consequently, outlier CVs were identified at the 1% probability level; assuming an approximately gaussian distribution, the individual estimate within these sets that was furthest from the observed mean of the set was removed, and this process was repeated until there were no more outliers at the 1% probability level. This process resulted in the deletion of 2.0% and 1.1% of the total original data for the WHO Standard and U.S. Standard, respectively.

The second outlier rule was applied to the logarithm of the ratio of U.S. mean mass to the WHO mean mass (within-collaborator, -run, -preparation, -dilution). Again, a 1% outlier rule was applied repetitively until no more outliers were detected. The result was an additional deletion of 3.2% and 3.4% of the total original data from the WHO Standard and U.S. Standard, respectively. Two additional procedures that were needed to delete remaining inconsistent or inappropriate data are discussed in Results.

Analysis of variance was used to quantitate the variability and to categorize the various sources of error of the relative potency and the mass concentration of the two preparations and the ratio for AFP mass/int. unit of the WHO Standard.

Results

Figure 1 illustrates a modified Youden plot (8, 9) of the results in mass concentration units for the U.S. Standard vs the 400 int. units/mL solution of the WHO Standard before any deletion of outliers; each point of inflection represents the mean of usually six determinations from three runs (\(n = 18\)) at each dilution multiplied by the dilution factor for each determination. The mean of the most concentrated AFP sample assayed by each collaborator is indicated by the accentuated point at one end of a series of interconnected lines;
the succession of inflection points each represents the mean results of the ordered series of increasing dilutions assayed by each collaborator.

In general, Youden plots graphically demonstrate the magnitude of analytical bias relative to random error. For pure random error, each of the four quadrants defined by the orthogonal medians should contain equal numbers of data points. Absence of any analytical error would result in a plot consisting of a single point. The analyzed results concentrated in the upper-right and lower-left quadrants in this study indicate that individual collaborators tended to get either high results with both materials or low results with both materials. In general, this fact is evidence for individual collaborator bias, which could be ascribed to differences in reagents or laboratory methodology used by each collaborator or, more probably, to differences in the "local" calibrators used. However, in addition to a fairly high among-collaborator mass concentration bias, this plot shows a very high within-collaborator, among-dilution bias for mass concentration estimates by some collaborators (B, K, I). For example, relative to his own mass concentration estimates, collaborator B tends to overestimate progressively the higher dilutions of both standards, so that when his estimates of the mass concentrations of his dilutions are multiplied by his respective dilution factors, a highly inconsistent mass concentration estimate is obtained from his estimates of either standard at their greatest concentration of AFP, 400 int. units/mL.

Note: Evaluator P.S. comments: As intended, use of the Youden plot reveals considerable among-laboratory bias when "local" mass units are used as well as considerable amplification of systematic variations related to any lack of parallelism with local calibrators. When similar variations are represented in a more conventional way (as in Figure 2), parallelism between the dose–response curves of the WHO Standard and individual calibrators was pretty good, if one excludes the most concentrated solution assayed by collaborator I and the least concentrated assayed by collaborator B.

From each collaborator's estimates of the AFP mass concentration for each measured dilution of the WHO Standard, we were able to estimate the number of nanograms equivalent to 1 int. unit of AFP. The individual means, usually based on six estimates per three runs (n = 18) of the mass/int. unit ratio of each dilution measured throughout the clinical range of interest, are plotted in Figure 2 for all collaborators after approximately 5% of the total data was deleted by the two outlier rules described in Methods. An acceptable AFP assay, used to screen maternal sera for the prenatal detection of neural-tube defects, must be able to give a uniform mass/int. unit ratio throughout most of the region of normal and pathological pregnancies.

For clinical orientation, Figure 2 also shows idealized log-normal distributions of serum AFP values expected for normal men (10) and for pregnant women at the 17th gestational week who are carrying normal, spina bifida, and anencephalic fetuses (derived from references 11 and 12 by assuming a normal median AFP value of 40 int. units/mL). These individual distributions have been normalized for equal areas, i.e., drawn for populations of equal size. The pregnancy distributions move from left to right by approximately 20% per gestational week during the mid-trimester. For neural-tube defect screening programs in the U.S., the relative magnitude of spina bifida and anencephalic populations are approximately equal, between 1 and 10 per 10,000 pregnancies, depending on race, geography, and possibly other genetic or environmental factors. Naturally, whatever cutoff value is chosen for interpretive (consultative) reporting of laboratory results, both false-negative and false-positive predictions of pregnancy outcome will occur in screening because of the large overlap of the distributions; the predictive value of a positive assay result for a particular pregnancy depends on accurate knowledge of the neural-tube defect prevalences for the different racial and geographic populations served (13, 14).

The fact that some collaborators' estimates of mass/int. unit (Figure 2) are completely above or below the group mean is further evidence for among-collaborator bias in mass determination. Internal consistency of individual collaborator's estimates of the ratio for mass/int. unit over the clinical range of concentrations evaluated is demonstrated by a relatively horizontal line connecting the individual estimates, such as obtained for collaborators C and E, and is a consequence of parallelism between (analytical response/log concentration)

---

**Fig. 2.** Nanograms equivalent to 1 int. unit of AFP from each collaborator's estimates of the AFP mass concentration for each dilution of the WHO Standard (72/225) over the region of clinical interest (indicated by the bell-shaped distribution curves).

Each point represents the mean of six estimates per three runs, i.e., n = 18.
plots of these individual collaborator's working curves and the WHO Standard. In contrast, collaborators B and K (both using reagents from the same manufacturer) obtained internally consistent results only between 400 and 100 int. units/mL, with varied values at lower AFP concentrations. In particular, collaborator B overestimates the AFP concentration in the WHO Standard (and in patients' sera) at 40 int. units/mL, the 17th week median AFP concentration for unaffected pregnancies, and underestimates AFP at 100 int. units/mL, a common 17th week cutoff value for pathological concentrations of AFP. When cutoff is expressed as a multiple of the unaffected population median, collaborator B will obtain a much smaller value than others for the same false-positive/false-negative ratio.

The mass/int. unit estimates from collaborator I were internally consistent for AFP concentrations appropriate for monitoring chemotherapy for germ-cell tumors (discounting the mean estimate at 20 int. units/mL, a commonly accepted cutoff value for distinguishing the presence of a tumor, as too high an AFP concentration for his range). However, all other estimates from I were biased high with respect to estimates of this ratio from other collaborators' AFP values appropriate for pregnancy sera.

Table 1 gives the CV values of the AFP mass/int. unit ratio estimated from individual collaborator's data at each dilution within the AFP concentration region of obstetrical interest. Ideally, the CV curve formed by connecting the values of any one collaborator (Table 1) should be flat or have a minimum near 100 int. units/mL, which is an AFP concentration near cutoff values (Figure 2) often chosen to minimize both false-positive and false-negative predictions of pregnancy outcome when screening for neural-tube defects (11, 12, 14). Analysis of variance of the 671 individual log(mass/int. unit) values obtained indicated that the principal sources of error were associated with: among-collaborator (53%), among-dilution (29%), among-replicate (15%), and among-run (3%) variability. An analysis of 30 slopes derived from individual collaborator's estimates after all outliers were deleted for each of the two standards for regressions of the log of reported AFP concentration in the region of obstetrical interest vs the log of the dilution measured showed that the two mean slopes for the two standards were parallel to each other within 5% and did not significantly differ from 1.00. If we included in the calculations only those values of the ratio mass/int. unit in the

<table>
<thead>
<tr>
<th>Collaborator</th>
<th>20</th>
<th>40</th>
<th>100</th>
<th>200</th>
<th>240</th>
<th>320</th>
<th>n</th>
<th>ng/int. unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8.6</td>
<td>8.6</td>
<td>8.3</td>
<td>8.8</td>
<td>8.9</td>
<td>10.5</td>
<td>15</td>
<td>0.982</td>
</tr>
<tr>
<td>B</td>
<td>4.6</td>
<td>7.6</td>
<td>12.6</td>
<td>11.6</td>
<td>8.0</td>
<td>8.0</td>
<td>12</td>
<td>0.945</td>
</tr>
<tr>
<td>C</td>
<td>4.6</td>
<td>4.0</td>
<td>2.8</td>
<td>2.7</td>
<td>2.8</td>
<td>5.0</td>
<td>15</td>
<td>0.884</td>
</tr>
<tr>
<td>D</td>
<td>8.0</td>
<td>7.2</td>
<td>5.6</td>
<td>4.4</td>
<td>4.0</td>
<td>4.2</td>
<td>15</td>
<td>1.271</td>
</tr>
<tr>
<td>E</td>
<td>11.8</td>
<td>11.7</td>
<td>8.0</td>
<td>6.8</td>
<td>7.3</td>
<td>7.3</td>
<td>15</td>
<td>0.887</td>
</tr>
<tr>
<td>F</td>
<td>10.2</td>
<td>10.5</td>
<td>11.5</td>
<td>12.2</td>
<td>12.2</td>
<td>11.8</td>
<td>15</td>
<td>1.120</td>
</tr>
<tr>
<td>G</td>
<td>3.7</td>
<td>4.1</td>
<td>4.1</td>
<td>4.1</td>
<td>4.1</td>
<td>4.1</td>
<td>6</td>
<td>1.300</td>
</tr>
<tr>
<td>H</td>
<td>6.0</td>
<td>6.0</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td>15</td>
<td>1.260</td>
</tr>
<tr>
<td>I</td>
<td>6.2</td>
<td>6.5</td>
<td>6.7</td>
<td>7.0</td>
<td>6.8</td>
<td>6.4</td>
<td>20</td>
<td>1.209</td>
</tr>
<tr>
<td>J</td>
<td>12.6</td>
<td>11.6</td>
<td>10.2</td>
<td>10.0</td>
<td>9.8</td>
<td>9.8</td>
<td>9</td>
<td>0.887</td>
</tr>
</tbody>
</table>

Overall, upper 95% confidence limit of mean: 1.19
Overall, n-weighted geometric mean: 1.06
Overall, lower 95% confidence limit of mean: 0.95

* Determined from plots of a second-degree polynomial regression fitted to each collaborator's within-run, within-dilution geometric coefficient of variation estimates.

b See text for outlier deletions.
collaborators B and I (Figure 2) that were above the 99th percentile of all collaborators' estimates of mass/int. unit. The clustering of data points around a 45° line through the intersection of the orthogonal medians of Figure 3 clearly indicates that, even after deletion of outliers and restriction of the concentration range under consideration, the mean relative potency of the U.S. Standard, as compared with the WHO Standard (i.e., the ratio of mass units each), is distinctly less variable than is the mean potency of either standard expressed in mass units alone.

We used this more uniform data set to assign an international unitage to the U.S. Standard in this study, based on the ratio between the reported mass values for equal dilutions of the U.S. Standard and the 400 int. units/mL solution of the WHO Standard. One estimate of relative potency (usually from the means of six mass determinations for both the numerator and the denominator) per dilution (usually five) per run (always three) was available from each of 10 collaborators. Analysis of variance of the 132 log(relative potency) values that remained after all the aforementioned outliers were deleted indicated that among-collaborator differences accounted for <1% of the overall variation in relative potency. For the average collaborator, the among-run and among-dilution components of error accounted for 37% and 62%, respectively, of the total variation of relative potency. The larger percentage for the dilution component may indicate insensitivity of some collaborators' assays at lower AFP concentrations, as well as some lack of exact parallelism between the dose-response curves given by the individual collaborators' calibrator (working curve) and the two standards of this study.

These results for relative potency, reflecting the use of the WHO Standard as a common calibrator (in international units), contrast greatly with those obtained in this study by use of "local" calibrators (in mass units). In the latter situation, more than half (53%) of the total variation was due to among-collaborator differences, in contrast to <1% obtained by use of a common calibrator.

Table 2 gives individual collaborator's estimates of the mean relative potency of the U.S. Standard, ranging between 0.937 and 1.039 of the 400 int. units/mL solution of the WHO Standard. The overall n-weighted geometric mean relative potency of the U.S. Standard relative to the 400 int. units/mL dilution of the WHO Standard was 0.982, thus providing an estimated mean AFP value for the U.S. Standard of 393 int. units/mL. Although the approximate 95% confidence interval for this mean estimate was ±10 int. units/mL (the CV of the mean was about 1.2%), as a result of this study we have assigned (defined) the AFP concentration of the U.S. Standard to be exactly 393 int. units/mL, when each vial is reconstituted precisely in accordance with the instructions accompanying the product. The among-vial variability of AFP content owing to imprecision of fill was <1%.

Interpretation of preliminary data on thermal instability of the U.S. Standard indicated that only the specimens held at 55 °C showed any statistically significant degradation when assayed during the first month by RIA and by immunofluorometric assay (data not presented); a half-life of about 5.5 months could be extrapolated from the curve for first-order decay at 55 °C. If we assume that this rate of decay would decrease twofold for each 10 °C drop in storage temperature, the calculated half-life at −20 °C, the usual storage temperature for the lyophilized standard is about 100 years. An Arrhenius analysis of "accelerated" thermal instability has been initiated, to extend over one year for the U.S. Standard; results will be reported separately, when completed.

**Discussion**

The principal reason for screening maternal serum for above-normal concentrations of AFP is for the prenatal detection of a fetus with open spina bifida. (Anencephalics invariably are stillborn or die shortly after birth, so do not present a continuing medical or survival problem for the offspring, the family, or society.) An ideal reference standard material would be a large pool of mid-pregnancy maternal sera (with known AFP concentration) obtained from women carrying a spina bifida fetus, because such sera would contain all the mid-pregnancy proteins associated with the pathology of interest; its collection, however, would be completely impractical and possibly unethical. The NCCLS AFP Subcommittee considered and rejected the idea of using maternal sera obtained late in the pregnancy or of enriching with AFP a pool of surplus mid-pregnancy maternal sera from unaffected pregnancies obtained from screening programs. Such pools would somewhat resemble sera from spina bifida pregnancies, in that they would contain trace amounts of many pregnancy-associated proteins other than AFP. However, this requirement was considered unnecessary because the primary purpose of the U.S. Standard is to calibrate AFP assays, not to act as a reagent to assess the immunochemical specificity of AFP tests, an independent problem. On the other hand, it was considered important to avoid the routine 250-fold or more dilution that would be required if the U.S. Standard for maternal serum AFP were a lyophilized specimen of pooled cord-blood serum.

In their 1974 study involving nine collaborators, Sizaret et al. (1) provided the basis for establishing the lyophilized pool of human placental cord serum (no. 72/225) as the WHO Standard on which the international unit for AFP is based (2). These authors concluded that "it is presently impossible to express 72/225 AFP activity in weight units" (1). In their 1979 re-evaluation of the AFP mass of the WHO standard, involving 11 collaborators, they concluded that "although the agreement between AFP experts is better now than it was a few years ago, we think that differences between individual estimates are still

<table>
<thead>
<tr>
<th>Collaborator</th>
<th>n *</th>
<th>Geometric mean ratio</th>
<th>Geometric CV, % b</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15</td>
<td>0.976</td>
<td>8.0</td>
</tr>
<tr>
<td>B</td>
<td>12</td>
<td>1.022</td>
<td>14.7</td>
</tr>
<tr>
<td>C</td>
<td>13</td>
<td>1.014</td>
<td>6.2</td>
</tr>
<tr>
<td>D</td>
<td>15</td>
<td>0.993</td>
<td>10.0</td>
</tr>
<tr>
<td>E</td>
<td>15</td>
<td>0.937</td>
<td>6.7</td>
</tr>
<tr>
<td>F</td>
<td>15</td>
<td>0.940</td>
<td>10.3</td>
</tr>
<tr>
<td>G</td>
<td>6</td>
<td>0.988</td>
<td>3.8</td>
</tr>
<tr>
<td>H</td>
<td>15</td>
<td>0.966</td>
<td>5.3</td>
</tr>
<tr>
<td>J</td>
<td>17</td>
<td>1.039</td>
<td>13.9</td>
</tr>
<tr>
<td>K</td>
<td>9</td>
<td>0.944</td>
<td>13.7</td>
</tr>
</tbody>
</table>

* See text for outlier deletions. b Includes among-run and among-dilution components. c International units per milliliter of the U.S. Standard when reconstituted with 0.50 mL of distilled water; values obtained by multiplying 400 int. units/mL by the relative potency ratios.
sufficiently important to justify the continuation of the use of the international unitage" (4). In the latter study, Sizaret gave the geometric mean and 90% confidence limits of the first study derived from the nine collaborators' estimates of the ratio for ng/int. unit of AFP. The geometric mean and 90% confidence intervals of the first study mean are plotted in Figure 4A (nine estimates of the first study were not individually reported). Figure 4B shows the 11 individual collaborator's estimates, geometric means, and 90% confidence intervals of the mean derived from various selections of the 11 estimates of the second study of Sizaret. They felt that their "best" estimate was 1.21 ng/int. unit (Figure 4, B6) because the six collaborators who contributed values for this estimate independently purified their own AFP. They also drew attention to the tight grouping of estimates from seven collaborators (Figure 4, B7), but could find no independent statistical justification for selecting this data subset. Our 10 collaborators' estimates (from Table 1), the geometric means, and 90% confidence intervals determined in this study are plotted in Figure 4C. Additional independent estimates of the ratio, nanograms of AFP per international unit, are plotted in Figure 4D: five estimates obtained in 1976 by Milford-Ward et al. (15), and two estimates in the same year by Keyser et al. (16); five estimates from commercially available kits obtained in 1979 by Kjaersgaard and Norgaard-Pedersen (17); two estimates of the ratio obtained in 1979 by Wong et al. (18) from their evaluation of commercially available kits; a 1980 estimate of the ratio by Griffiths et al. (19) from assessment of the Canadian Provisional Reference Standard for AFP; and nine estimates of the ratio by manufacturers of the kits used in this study, obtained before the start of this study (personal communications). The overall geometric mean of these 45 "independent" determinations of the ratio is 1.126 ng/int. unit, the 90% confidence intervals of this mean is 1.083 and 1.171 ng/int. unit, and the median is 1.120 ng/int. unit. Note that although this overall geometric mean is included within the 90% confidence intervals of the means experimentally determined by Sizaret et al. and by us, the "best" estimate of Sizaret (4), 1.21 ng/int. unit (Figure 4, B6) and our "best" estimate, 1.06 ng/int. unit (Figure 4, C10), are both outside the 90% confidence intervals of the mean estimate obtained from the 45 individual estimates (Figure 4, E45).

Note: Evaluator P.S. comments: The so-called 45 "independent" determinations were independent in the sense that they were made by different people, but they were probably not independent chemically, since some of the laboratories probably were using calibrators of the same origin. The spread of the population of values does not necessarily increase or decrease with the number of estimates. If each estimate is properly a member of the same population of estimates, then the mean estimate (standard error of the mean) is improved by increasing the number of estimates.

We have reviewed the variability of previous estimates of the ratio of nanograms per international unit of AFP summarized by Figure 4, and we have attempted in this study to improve estimates of this ratio, with the cooperation of U.S. experts, particularly those who probably will supply most of the working curve "standards" used in the U.S. In light of these considerations, we fully concur with the 1979 conclusions of Sizaret (4) that interlaboratory variability for AFP concentration is lessened if all results are reported in international units after appropriate calibration of "local" working curve preparations against the AFP content of the WHO standard (no. 72/225) or its national equivalent, wherein AFP concentration is expressed in int. units/mL. This is also the current position of the United Kingdom Department of Health and Social Security Supraregional Assay Service for England, Wales, and Northern Ireland for its National External Quality Assessment Scheme for Maternal Serum Alpha-Fetoprotein (20). To obtain numerical agreement among the latter laboratories in the U.K. quality-assurance program, a precondition for participation is that all AFP values must be reported in international units derived from the First British Standard for Human Cord Serum (no. 72/227), which was produced from the same batch and cooperatively calibrated in international units against the WHO Standard (no. 72/225).

Nevertheless, having given due consideration to recommendations for the use of international units rather than mass units for AFP standardization of clinical assays, we believe that some research applications for the U.S. Standard may necessitate assignment of mass units to it. Consequently, we have assigned the value 442 ng/mL (392.9 int. units/mL X 1.126 ng/int. unit) as the most nearly accurate current consensus estimate of the "true" value of AFP mass concentration in each vial of the U.S. Standard when it is reconstituted with 0.50 mL of distilled water according to the instructions accompanying the product. This consensus approach for establishing the "true" concentration of an analyte in a reference preparation has proved useful for other protein analytes in a biological matrix (21).

Sizaret et al. (1, 4) have discussed some of the reasons why locally purified AFP used as a primary reference standard for local reference preparations may cause variable results:

These large differences between local standards may correspond to [a] unequal purity of local standards, some of them being possibly contaminated by [other] proteins, acrylamide, chemicals, water, etc.; these last three causes of contamination can be taken into account only when AFP activity has been estimated by weighing; [b] lack of convergence of techniques used for estimating AFP activity (weights and optical densities); [c] unequal activity of purified AFP's due to unequal antigenticities . . . . [d] AFP microheterogeneity . . . . [and e] use of nonspecific antisera; of the nine (antisera) of Sizaret's first study (1), only five antisera were found to be monospecific.

In addition, the process of AFP purification can degrade and cause aggregation of native AFP, which may influence immunoassay results (22). The report of the NCCLS AFP Subcommittee (5) points out that: "no pure standard for na-
tive AFP, collaboratively calibrated in mass units, currently exists. Properly influenced by consideration of the problems noted above as well as by cost, the Committee did not recommend development of a pure AFP national or international reference preparation. This same conclusion had been reached in 1978 by a U.S. National Institutes of Health AFP working group (23), in their recommendation for development of a national reference preparation for AFP in a biological matrix, calibrated in international units against the WHO Standard (no. 72/225).

Imunoassays are often performed for analytes in relatively low concentration in a complex molecular milieu. Some of the analytical variability caused by use of local calibrators may result from disregard of the analytical consequences of using different biological matrices in the unknowns and the local standards. Sazaret et al. (1, 3, 10) in several extensive investigations have demonstrated parallelism between curves relating analytical response to log(relative AFP concentration) for patients' sera and the WHO Standard. Accordingly, we have not reinvestigated the necessary condition of analytical parallelism between the U.S. Standard and clinical specimens in this study, because the WHO Standard and the U.S. Standard demonstrated appropriate parallelism with each other.

The U.S. National Reference Preparation for Alpha-Fetoprotein in Mid-Pregnancy Maternal Serum may be obtained free of charge by writing to the Chief, Immunological Products Branch, Centers for Disease Control, Atlanta, GA 30333. To preserve an adequate supply of the U.S. Standard for the next 10 years, yet ensure that it is readily available to those who most need it, restrictions on its distribution have been set. Centers for maternal serum AFP screening that attest in their letters of request, signed by the laboratory director, that their center currently is engaged in, or in the near future, will be screening for neural-tube defects with a workload of at least 50 patient specimens per week (5, 23), and manufacturers of AFP reagents who contemplate serving such centers, may obtain six vials of the U.S. Standard free of charge every six months; all others may have one vial per year.

Comments

Evaluator J.E.H.: One of the difficulties experienced by investigators analyzing AFP by radioimmunoassay has been the lack of a readily obtainable common standard; the WHO Standard is not readily and frequently available to individual laboratories in the United States. Although it is possible to compare results for maternal serum AFP between laboratories via the multiples of the median (MOM) method, such comparisons can be simplified by a commonly accepted unit. More importantly, availability of a common calibrator can facilitate proficiency testing programs. Towards this end, the CDC has carefully prepared a U.S. AFP Standard of defined potency, calibrated against the WHO standard, dispensed in ampoules with enough material for a single calibration and made available to individual laboratories. This represents a significant step forward in quality control in this field.

Evaluator W.M.H.: This paper reports an important initiative in the implementation of the sound WHO policy, which is already widely accepted in the U.K., of providing standards for "biological" materials, i.e., those whose structure is not fully defined and therefore not synthesized in the form of sealed ampoules for which activity is designated after extensive interlaboratory evaluation. The advantages of providing a consistent common currency, not only within the U.S. among users of the new U.S. Standard, but internationally through its tying to the WHO Standard, include the minimization of errors that may occur among clinicians and laboratory staff who may be involved with data from different centers. In addition, temporal changes due to unrecognized potency shifts in stored local standards are removed. The problems of external quality assessment of AFP assays will be reduced by avoiding the systematic and arithmetic errors associated with applying conversion factors. The exclusion of variability of calibrant as a source of between-laboratory variability allows consideration of other problems, e.g., the discovery that assays may be biased because they use an inappropriate matrix for their standards or have a high matrix variability. Active external quality assessment as practiced in the U.K. would be almost impossible without the adoption of a common standard.

The decision to have the AFP activity pitched at the one assay per ampoule level has a further advantage: it reduces the danger that users may use stored diluted standard, a practice shown to be error prone (Hunter, W. M., McKenzie, L., and Bacon, R. R. A., In Immunoassays for the 80s. A. Voller, A. Bartlett, and D. Bidwell, Eds., MTP Press Ltd., 1981, pp 155-165). There are purposes for which units of mass are necessarily used, e.g., for comparison of cross reactions among homologous protein series, and these justify the adoption of consensus "mass potency" for the proposed standard. However, there is a danger that this currency may be adopted for other, more routine purposes. The paper should be stronger in advising against this and should spell out the dangers of so doing. I am particularly troubled by the decision to put mass units on the ampoule label. This could be a big mistake.

Submitter C.B.R.: One might take a position that whatever unitage was used is of little consequence, provided all laboratories used a common calibrator. The problem with this is that when mass concentration units are used there is no intrinsic assurance that they are based on the mass concentration assigned to the common calibrator. If they are actually based on local mass units, the among-laboratory bias demonstrated by Figures 1 or 4 will apply. Only when WHO international units are used can one be reasonably assured that the calibrator values are referable to a commonly used reference preparation. The CDC favors use of WHO International units. Vials of the U.S. Standard were labeled before this manuscript was written—prominently in int. units/mL, and parenthetically in ng/mL.

We thank Drs. James M. Barbaree, Biological Products Division, CDC, for freeze-drying and labeling the vials of the U.S. Standard, and Joan C. May, Division of Biochemistry and Biophysics, Bureau of Biologics, FDA, for analyzing the residual moisture in the freeze-dried U.S. Standard. We are particularly grateful to the following persons and organizations who collaborated with CDC to provide the AFP assays that constitute the basis for this report. The assignment of code designations to these collaborators does not relate to their order of listing here.

Lawrence M. Cummins, Ph.D.
Abbott Diagnostics
Chicago, IL 60064

Amiram Daniel, Ph.D.
Bureau of Medical Devices Laboratory
Food and Drug Administration
Washington, DC 20250

Phillip G. Douglas, Ph.D.
Amersham Corporation
Arlington Heights, IL 60005

Herbert A. Fritcher, Jr., Ph.D.
M.D. Anderson Hospital
Houston, TX 77025

Mr. Bertram W. Griffiths
Bureau of Medical Biochemistry
Laboratory Centre for Disease Control
Ottawa, Ontario K1A0L2
Canada

George Knight, Ph.D.
Foundation for Blood Research
Scarborough, ME 04074


---

**Editor's note:** The reader is reminded that Selected Methods do not bear the official imprimatur of the Association. They are methods that seem durable and generally useful, and that have been checked by several evaluators. As detailed elsewhere [Clin. Chem. 19, 1207 (1973)], these methods are offered here for criticism by the world community of users, and will be revised appropriately before being collected into a bound volume, Selected Methods of Clinical Chemistry. The last such volume was published by the Association in 1977.