Anal. Chem. **2000,** *72,* 4163−4170

NIST Micronutrients Measurement Quality Assurance Program: Characterizing the Measurement Community's Performance over Time

David L. Duewer,* Margaret C. Kline, Katherine E. Sharpless, and Jeanice Brown Thomas

Chemical Science and Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8394

The Micronutrients Measurement Quality Assurance Program (M^2QAP) at the National Institute of Standards and Technology was created in 1984 with the goal of improving among-participant measurement comparability for fat-soluble vitamin-related compounds in human serum. We recently described improved tools for evaluating comparison exercise data; we here extend and apply these tools to the evaluation of the measurement community's performance over the entire 15-year history of the M^2QAP. We here display measurement performance characteristics for the 14 measurands most commonly reported by the M^2QAP community. We confirm that *among*-participant comparability for total *β*-carotene cannot be much improved without improving average long-term *within*-participant measurement stability. We demonstrate that improved measurand definition and/or identification of interferences may help participants improve comparability for many of the M^2QAP's other commonly reported measurands. The reported measurement performance characteristics may be of interest to clinical, nutritional, and epidemiological studies involving any of these measurands. The data analysis techniques utilized may be applicable to other programs.

The National Institute of Standards and Technology (NIST) and the National Cancer Institute (NCI) established the Micronutrients Measurement Quality Assurance Program (M^2QAP) in 1984. The original focus of the M^2QAP was development of the measurement infrastructure required for long-term studies involving *β*-carotene, α-tocopherol, and/or retinol in human serum.[1] Over the course of 15 years, many additional fat-soluble, vitamin-related compounds have become of clinical, nutritional, and epidemiological interest. Fourteen analytes are now routinely reported by a sizable fraction of comparison exercise participants: total α-carotene, *trans*-*β*-carotene, total *cis*-*β*-carotene, total *β*-carotene, total *β*-cryptoxanthin, total lycopene, *trans*-lycopene, total lutein, total zeaxanthin, total lutein plus zeaxanthin, total and/or *trans*-retinol, total retinyl palmitate, α-tocopherol, and *γ*-tocopherol.

While no longer supported by NCI for epidemiological and/or intervention studies, the primary goal of the M^2QAP is unchanged: enabling valid comparison of relevant measurements over time and among laboratories.[2] Since participation is voluntary, measurement comparability can only be enhanced by providing participants with materials and information that help them improve their own measurement processes. To this end, we recently introduced conceptual and graphical analysis tools intended to help participants better interpret their results.[3,4] The most useful of these tools is the quantitative dissection of among-participant measurement comparability into two components: (1) concordance, the average of the differences between an individual participant's measurements and the consensus results for a given set of samples, and (2) apparent precision, the standard deviation of those differences.

In a previous study of total measurement variance, we concluded that improved within-participant quality control was required to achieve improved among-participant measurement comparability.[1] Analysis of the concordance and apparent precision characteristics of each of the analytes routinely reported in the M^2QAP suggests that additional strategies are needed. The most critical means to improve comparability may sometimes be the explicit and unambiguous definition of what chemical entity(ies) are specified as a given named analyte.

We present here the measurement performance characteristics for all routinely reported M^2QAP fat-soluble, vitamin-related analytes. The compound-specific details are primarily of interest to the clinical and nutritional vitamin measurement community and to any epidemiological or meta-analytic study using their measurements. The data analysis techniques illustrated for these analytes are appropriate for use by many measurement communities interested in improving their ability to accurately determine important compounds and to validly exchange analytical results among laboratories.

* Corresponding author: David Lee Duewer, NIST, 100 Bureau Drive, Stop 8394, Gaithersburg, MD 20899-8394. Tel.: 301-975-3935. Fax: 301-977-0685. E-mail: david.duewer@nist.gov.

(1) Brown Thomas, J., Sharpless, K. E., Eds. *Methods for Analysis of Cancer Chemopreventive Agents in Human Serum*. Special Publication 874; National Institute of Standards and Technology, US Government Printing Office: Washington, D.C., 1995.

(2) Duewer, D. L.; Brown Thomas, J.; Kline, M. C.; MacCrehan, W. A.; Schaffer, R.; Sharpless, K. E.; May, W. E.; Crowell, J. A. *Anal. Chem*. **1997**, *69*, 1406−1413.

(3) Duewer, D. L.; Kline, M. C.; Sharpless, K. E.; Brown Thomas, J.; Gary, K. T.; Sowell, A. L. *Anal. Chem*. **1999**, *71*, 1870−1878.

(4) Duewer, D. L.; Kline, M. C.; Sharpless, K. E.; Brown Thomas, J.; Stacewicz-Sapuntzakis, M.; Sowell, A. L. *Anal. Chem*. **2000**, in press.

10.1021/ac991480j
Published on Web 08/02/2000

MATERIALS AND METHODS

Materials. All measurements considered in this study are routinely reported results from M²QAP participants for the analysis of human serum samples. All samples have been prepared at NIST, nearly all as lyophilized sera. The sera used in M²QAP comparisons have been obtained and processed in accordance with U. S. federal standards for handling blood-derived materials.

Analytical Methods. M²QAP participants have diverse measurement needs and resources.[2] While most involve some form of sample extraction, solvent-exchange, and reversed-phase liquid chromatography, the methods employed represent widely different compromises among sample size, analysis time, sensitivity, and selectivity.[1]

Measurands. Of the 14 commonly reported "analytes", only α-tocopherol, γ-tocopherol, *trans*-β-carotene, and *trans*-lycopene are single geometrical isomers. The "total" analytes explicitly group the *trans* and *cis* isomers as a single measured quantity ("measurand"), generally derived from the area of a single chromatographic peak or retention interval. However, some participants do separately report results for *trans*- and *cis*-β-carotene isomers: the M²QAP policy is to sum these values into a "total" result (if the participants have not done so themselves.) Likewise, the difference between separately reported total and *trans*- values is recorded as "total *cis*-". All of the geometric isomers of lutein and zeaxanthin are often reported as the single measurand "total lutein and zeaxanthin"; when total lutein and total zeaxanthin values are reported, their sum is also recorded as total lutein and zeaxanthin.

While *trans*- and *cis*-isomers of retinol exist and have different biological activities,[5] there was no M²QAP policy on differentiating retinol isomers throughout the period reported in this study. The measurand "retinol" thus designates *trans*-retinol, plus a variable mix of *cis*-retinol isomers specific to each participant's chromatographic method and reporting policy.

Data. This study uses essentially all measurements reported for any of the above 14 measurands for any of 172 homogeneous samples distributed during the 42 M²QAP comparisons conducted from late 1984 through mid 1999. Only the following data are excluded: (1) qualitative values ("0", "not detected", and "detected but not quantified"), (2) data for any analyte not reported by at least five participants in a given exercise, and (3) data for samples determined by NIST analysts to be heterogeneous. All semiquantitative values (those reported as upper or lower bounds) have been treated as having the value of the bound.

DEFINITIONS

The nomenclature and calculations developed for characterizing among-participant measurement performance are presented in detail elsewhere.[3] Briefly, the M²QAP defines the standardized measurement comparability for a particular measurand X in a given sample *j* distributed in a given comparison *k* reported by a given participant *l* as the measurement *Z* score [6]

$$\Delta_{jkl} = \frac{X_{jkl} - \bar{X}_{jk}}{S(\bar{X}_{jk})} \tag{1}$$

where X_{jkl} is the measured concentration, $\bar{X}_{jk}$ is the consensus concentration, and $S(\bar{X}_{jk})$ is a fitted among-participant measurement standard deviation for the consensus concentration. The characteristic measurement concordance is the average of each participant's standardized comparability for every sample distributed in the exercise

$$C_{kl} = \sum_{j=1}^{N_j} \Delta_{jkl} / N_j \tag{2}$$

where N_j is the number of samples analyzed. The characteristic apparent precision is the standard deviation of the comparabilities

$$\mathrm{AP}_{kl} = \sqrt{\sum_{j=1}^{N_j} (\Delta_{jkl} - C_{kl})^2 / (N_j - 1)} \tag{3}$$

The expected deviation from consensus for a participant's measurement process is defined as the root sum of squared comparability and apparent precision

$$D_{kl} = \sqrt{\mathrm{C}_{kl}^2 + \mathrm{AP}_{kl}^2} \tag{4}$$

As discussed in detail elsewhere,[2] the conceptual model for all M²QAP measurands is that the reported values are drawn from two measurement populations: (1) a majority population that is roughly normal about the consensus value and (2) a minority population that is roughly uniform from "not detectable" through "way too high". We therefore use robust quartile-based statistics to estimate the consensus value, $\bar{X}_{jk}$, and among-participant standard deviation, S_{jk}: the 2nd quartile (median) and 0.741(3rd quartile − 1st quartile), respectively.[7] These estimates are insensitive to "outlier" data (up to just less than 50% of the values) as long as the numbers of too-small and too-large values are about equal.

While the robust S_{jk} estimate is appropriate when analyzing results for a particular exercise, it does not enable characterizing changes in measurement variability over time. The among-participant measurement standard deviation expected for a given concentration at some given period of time can be defined from regression of the empirical relationship

$$S(X) = \sqrt{L_{qc}^2 + (\beta_0 X^{\beta_1})^2} \tag{5}$$

on a particular set of observed ($\bar{X}_{kl}$, S_{jk}) pairs.[2] For this study, we parametrized L_{qc}, β_0, and β_1 for each of the 14 measurands using just the samples distributed in the 10 M²QAP exercises conducted from February 1996 through April 1999. Measurement comparability, concordance, apparent precision, and expected deviation thus have units of "expected among-participant measurement

(5) Weiser, H.; Somorjai, G. *Int. J. Vitam. Nutr. Res.* **1992**, *62*, 201−208.

(6) ASTM Committee E-36. E 1301-95: Standard Guide for Proficiency Testing by Interlaboratory Comparisons. *Annual Book of ASTM Standards, Vol. 14.02*; American Society for Testing Materials: West Conshohocken, PA, 1997; pp 809-822.

(7) Analytical Methods Committee, Royal Society of Chemistry. *Analyst* **1989**, *114*, 1693−1697.

standard deviation characteristic of the time period 1996 through mid 1999" or "$SD_{96:99}$".

RESULTS AND DISCUSSION

While most M²QAP samples are prepared in sufficient quantity for distribution in more than one exercise, only a few have been distributed in more than three exercises and none in more than five. In the absence of any "standard set" of samples reanalyzed at regular intervals, we seek to evaluate measurement performance characteristics over time using standardized measurement comparabilities (eq 1). Comparabilities for all samples should have the same distribution, within the limitations of the robust statistical estimates and the reality of our models for the measurement populations. Given that the goal of the M²QAP is to promote smaller among-participant measurement standard deviations over time, we must also be able to successfully predict the standard deviation expected for a given concentration for some given reference period (eq 5). Current M²QAP participants are naturally most concerned about their current measurement performance, hence we choose the recent past as our reference time period. (We define "recent" as the last 10 comparisons—currently March 1996 through March 1999. This is not completely arbitrary; since four samples are typically distributed in each exercise, 10 exercises provides a usefully large sample population.[3])

In the following sections we focus first upon the measurements reported during the 10 most recent comparisons, conducted from 1996 through mid 1999. We show that the "roughly normal majority" of the measurement performance characteristics calculated (eqs 2, 3, and 4) from these measurements are well summarized by median values. We then use the median performance characteristics for each of the 42 M²QAP exercises to summarize measurement performance over time.

Measurement Concordance Distributions. The empirical cumulative distribution or "ogive" of a set of values is visualized by plotting the rank-ordered values, sorted smallest to largest, against an estimate of the cumulative probability of observing the value. While variously defined,[8] we estimate the cumulative probability as

$$\mathrm{PSRank}_i = \mathrm{Rank}_i / (N + 1) \qquad (6)$$

where N is the size of the set and Rank_i is the ordinal rank (1 through N_i) of each element of the ordered set. The ogive is typically drawn using a probability axis transformed such that a set from a truly normal distribution will plot as a straight line.[9] Since our distributions are demonstrably nonnormal, a linear zero-to-one probability axis provides a more readily interpreted display for comparing and modeling multiple distributions. Figure 1 displays the comparability ogives for the 1996 through mid 1999 measurements of all 14 measurands and for the standardized normal distribution, generated using the inverse function

$$z = \Phi^{-1}(\mathrm{PSRank}, \mu = 0, \sigma = 1) \qquad (7)$$

where Φ is the normal (Gaussian) distribution, μ is the distribution mean, σ is the standard deviation, and z is density at probability PSRank.

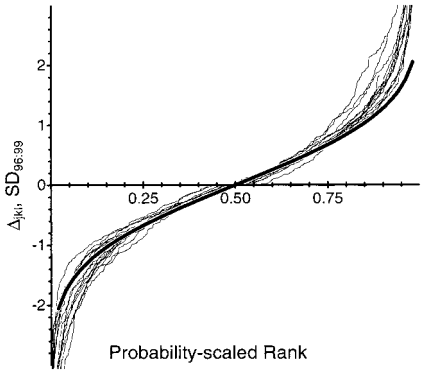

Figure 1. Comparability, Δ_{jkl}, as functions of probability-scaled rank order. Each of the 14 light lines represents the empirical ogive (cumulative distribution) for the Δ_{jkl} associated with every measurement reported for one measurand during the time period 1996 through mid 1999. The dark line is the ogive of the standardized normal (zero mean, unit standard deviation).

The ogives of all 14 measurands are quite similar over more than the central 50% of values (PSRank from 0.25 to 0.75). There is an excess of large-positive Δ_{jkl} (measurements larger than consensus) over large-negative (smaller than consensus), manifest as more curve density "above" the standardized normal at the right side of the plot than "below" it on the left. This reflects the larger concentration span between consensus and "way too high" than between consensus and "not detectable." All of the ogives display this behavior, regardless of the number of qualitative at-or-below quantitative limit data reported for the measurand; exclusion of these qualitative data in the definition of $\bar{X}_{jk}$ and S_{jk} thus does not account for the asymmetry. Rather, the asymmetric tails indicate that the measurements are better described as spanning a logarithmic range. This is consistent with the log-normal distributions observed for control-sample measurements made over a six-month period in our laboratory.[10]

While logarithmic transformation of the measurements removes the asymmetry of the distribution tails, it also complicates the communication of resulting tales. We note that more than the central 50% of the (nontransformed) Δ_{jkl} do align well with the standardized normal. The "roughly normal" measurements appear to be in sufficient majority and the "roughly uniform" minority to be sufficiently symmetrical for successful quartile-based estimation.

Comparability, Apparent Precision, and Expected Deviation Distributions. While the signed-concordance is necessary when interpreting a given participant's measurement performance,[4] the median signed concordance is, by definition, about zero. The absolute concordance, $|C_{kl}|$, is more appropriate for describing the expected magnitude of the measurement discordance among the measurement community.

The ogives of the three measurement performance characteristics are of identical form for all 14 measurands. Figure 2 illustrates these forms for total β-carotene, along with the following distribution models parametrized to the "roughly normal

(8) Kimball, B. F. *J. Am. Stat. Assoc.* **1960**, *55*, 546−560.
(9) Schmid, C. F., Schmid, S. E. *Handbook of Graphic Presentation*, 2nd ed.; Wiley: New York, NY, 1979.
(10) Sharpless, K. E.; Duewer, D. L. *Anal. Chem.* **1995**, *67*, 4416−4422.

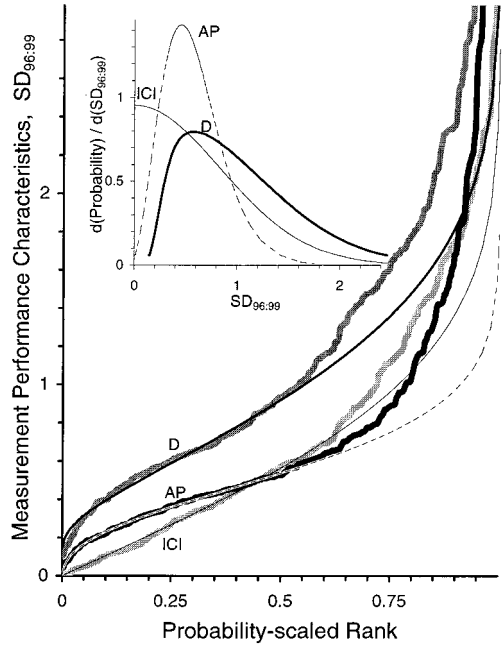

Figure 2. Absolute concordance, apparent precision, and expected deviation for total β-carotene as functions of probability-scaled rank order. The wide lines represent the empirical cumulative distributions for the three performance characteristics. The thin lines are the ogives for the parametrized model distributions (eqs 10a, 10b, and 10c). The inset displays the model distributions as probability density functions.

Table 1. Distribution Parameters for $|C|$ and AP[a]

		$\|C_{kl}\|$		AP_{kl}			D_{kl}	
measurand	N_p[b]	β_2[c]	± Rs[d]	β_3[e]	β_4[e]	± Rs[d]	ϵ_5[f]	± Rs[d]
total or *trans*-retinol	503	0.91	0.02	0.68	0.21	0.01	0.13	0.02
α-tocopherol	484	0.85	0.01	0.64	0.19	0.01	0.14	0.01
total β-carotene	338	0.84	0.01	0.73	0.20	0.01	0.14	0.02
total lycopene	283	1.01	0.02	0.57	0.15	0.01	0.10	0.02
total α-carotene	274	0.80	0.01	0.74	0.23	0.01	0.17	0.03
γ-tocopherol	261	0.91	0.01	0.66	0.17	0.01	0.12	0.03
total β-cryptoxanthin	250	0.93	0.01	0.58	0.18	0.01	0.13	0.02
total lutein and zeaxanthin	239	0.93	0.02	0.66	0.20	0.01	0.14	0.03
total lutein	157	0.80	0.02	0.71	0.22	0.02	0.15	0.03
trans-β-carotene	140	0.84	0.02	0.83	0.25	0.01	0.17	0.03
retinyl palmitate	139	0.84	0.02	0.81	0.27	0.01	0.19	0.03
total zeaxanthin	132	0.90	0.02	0.69	0.23	0.02	0.12	0.02
total cis-β-carotene	98	1.07	0.05	0.73	0.24	0.01	0.13	0.03
trans-lycopene	96	0.94	0.05	0.61	0.18	0.02	0.12	0.09

[a] Parameters have units of $SD_{96:99}$. [b] The number of $|C|$, AP, and D; the smallest magnitude 50% of these values was used to parametrize the model distributions. [c] Parameter of model distribution, eq 8a: $|C_{kl}|^* = \Phi^{-1}(((1 + \mathrm{PSRank}_{kl})/2),0.0,\beta_2)$ [d] Expected residual, $(\sum_{i=1}^{N_p/2}(Y_{\mathrm{calc}_i} - Y)^2_{\mathrm{obs}_i}/(N_p/2 - \text{no. parameters}))^{1/2}$ [e] Parameters for model distribution, eq 8b: $AP^*_{kl} = (\Phi^{-1}(\mathrm{PSRank}_{kl},\beta_3,\beta_4))^2$ [f] Parameter of model distribution, eq 8c: $D_{kl} = ((|C_{kl}|^*)^2 + (A^*_{kl})^2)^{1/2} + \epsilon_5$.

majority" 50% of the data:

$$|C_{kl}|^* = |\Phi^{-1}\left(\frac{1 + \mathrm{PSRank}_i}{2},0,\beta_2\right)| \tag{8a}$$

$$\mathrm{AP}^*_{kl} = (\Phi^{-1}(\mathrm{PSRank}_i,\beta_3,\beta_4))^2 \tag{8b}$$

$$D^*_{kl} = \sqrt{(|C_{kl}|^*)^2 + (\mathrm{AP}^*_{kl})^2} + \epsilon_5 \tag{8c}$$

While other models may describe the data equally well, these align adequately with the observed values over more than 50% of the data while using just four adjustable parameters for each measurand. The inset in Figure 2 displays these model distributions in the more familiar probability density function form, approximated as the first difference of PSRank with respect to the ordered values. Table 1 lists parameter and diagnostic values for the three model distributions for all 14 measurands.

To the extent that the models of eq 8 do describe "true" behavior, several of the parameters are readily interpretable. Parameter β_2 estimates the true median absolute concordance for the majority population

$$\mathrm{median}(|C_{kl}|) = |\Phi^{-1}\left(\frac{1 + 0.5}{2},0.0,\beta_2\right)| = \Phi^{-1}(0.75,0.0,\beta_2) \tag{9a}$$

Parameter β_3 likewise estimates the true median apparent precision

$$\mathrm{median}(\mathrm{AP}_{kl}) = (\Phi^{-1}(0.5,\beta_3,\beta_4))^2 = \beta_3^2 \tag{9b}$$

The observed medians and those predicted from the model parametrizations are in excellent accord for all 14 measurands.

Parameter ϵ_5 of eq 8c is truly empirical; this model was suggested by the data-driven insight of extremely few D_{kl} smaller than 0.1. Nearly all the parametrizations of eq 8c align as well or better to the smallest magnitude 50% of the data than do models having the form of eq 8b. (Alignment is inferior only for *trans*-lycopene, the measurand with fewest data.) More significantly, the eq 8c parametrizations are all more predictive of the "not so normal" largest magnitude 50%. Given that the values for ϵ_5 for all measurands are within a narrow range (0.10−0.19), we interpret this parameter as defining the smallest achievable deviation: 0.1−0.2 $SD_{96:99}$. We speculate that this value may be related to the average intrinsic sample heterogeneity.

Changes over Time: Original Measurands. Figure 3 displays the median absolute concordance, apparent precision, and the expected deviation for retinol, α-tocopherol, and total β-carotene for each of the 42 M^2QAP comparisons. The majority of participants have reported these three measurands from the very first exercise in 1984. The measurands have common trends: (1) With the exception of the very earliest exercises, apparent precision has remained within the range 0.3 to 0.8 × $SD_{96:99}$. (2) The expected deviation improved from the original 2 × $SD_{96:99}$ in 1984 to about 0.8 × $SD_{96:99}$ by 1989, remained at this minimum briefly then steadily rose to a bit more than the current level, reached a maximum during 1994 to 1996, and is now either stable or slowly improving. (3) Discordance was the dominant source of variance from 1984 to 1989, although concordance steadily improved during this period. Discordance and apparent imprecision contribute about equally to the total β-carotene variance from 1989 through mid 1999; both concordance and apparent precision

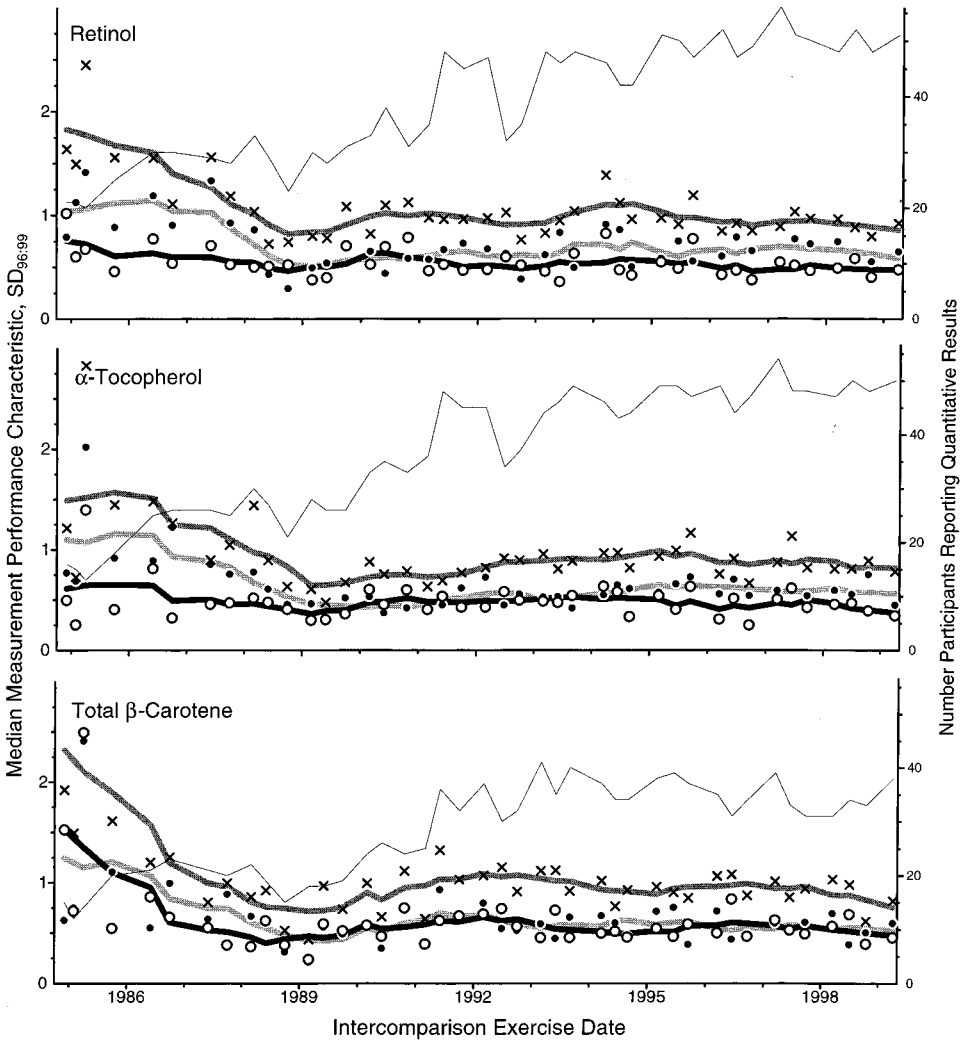

Figure 3. Measurement performance characteristics as a function of time for retinol, α-tocopherol, and total β-carotene. Each of the three segments displays the absolute concordance (solid circles, thick light gray line), apparent precision (open circles, thick solid line), and expected deviation ("x", thick dark gray line) estimated for the M^2QAP measurement community for all 42 comparisons conducted from late 1984 through mid 1999. The thick lines have been linearly smoothed using the 5 closest-in-time exercises. The light solid line denotes the number of participants reporting quantitative data for the measurand in each exercise.

have slowly improved from the secondary maximum reached in 1992. Discordance and apparent imprecision were also approximately equal for retinol and α-tocopherol in the early 1990s. While the magnitude of discordance for both measurands has remained roughly constant since 1994, apparent precision has slightly but consistently improved during this time so that discordance is again dominant.

Particularly for total β-carotene, the steady measurement performance improvement from 1984 to 1989 can be attributed to participation in the M^2QAP exercises. However, the 1989 to 1990 "best performance" period corresponds with a period in which sets of control materials were freely distributed to all participants. We believe that these materials were not only used to optimize (long-term) performance of many measurement systems but that some participants used these materials to calibrate their systems and thus "artificially" improve their (short-term) performance in the comparisons.

A factor in the general agreement among participants after 1989 may be the development and availability of the first version of NIST's Standard Reference Material (SRM) 968, Fat-Soluble Vitamins in Human Serum.[11] A primary role for this and for all SRMs is to help establish and maintain measurement agreement among laboratories. SRM 968 is an integral part of the M^2QAP: all four releases of SRM 968 have been routinely distributed as samples in one or more exercises.

While periodic analysis of SRM 968 by many participants may indeed help maintain among-participant concordance, such analyses are unlikely to improve apparent precision. Natural matrix SRMs are not only too expensive to use as routine control materials, their matrixes—however complex—are static and so pose only a fixed set of measurement challenges. Comparison exercise samples provide a much greater range of learning opportunities. Starting about 1991, the M^2QAP samples have been intentionally made more complex by using multiple serum pools, augmented measurand concentrations, and atypical relative con-

(11) National Institute of Standards and Technology. Certificates of Analysis: (1) Standard Reference Material 968, Fat-Soluble Vitamins in Human Serum, 1989. (2) SRM 968a, Fat-Soluble Vitamins in Human Serum, 1993. (3) SRM 968b, Fat-Soluble Vitamins and Cholesterol in Human Serum, 1995. (4) SRM 968c, Fat-Soluble Vitamins, Carotenoids, and Cholesterol in Human Serum, 1999. NIST Standard Reference Materials Program: Gaithersburg, MD

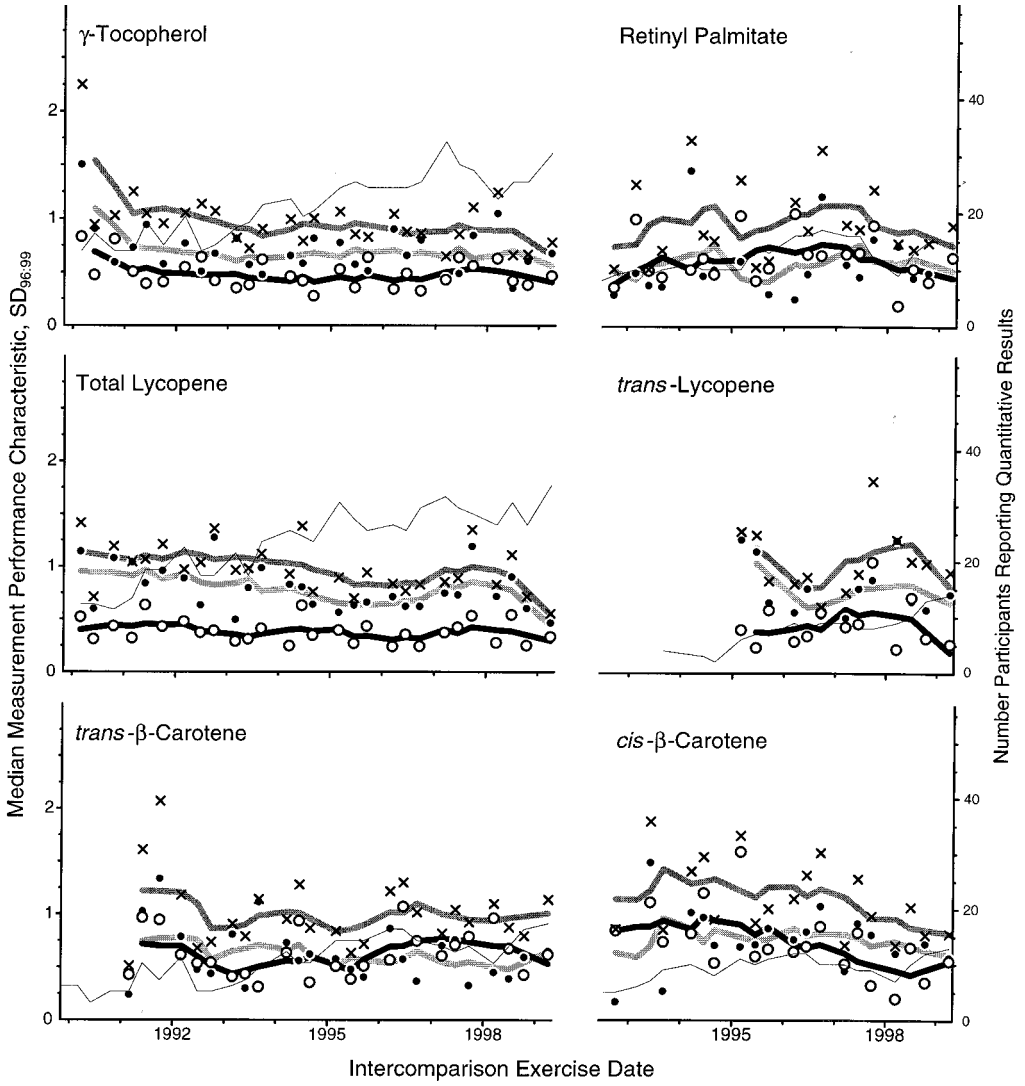

Figure 4. Measurement performance characteristics as a function of time for γ-tocopherol, retinyl palmitate, total lycopene, *trans*-lycopene, trans-β-carotene, and cis-β-carotene. Legend as in Figure 3.

centrations of some measurands. The measurement challenges designed into each exercise also became more varied as a greater variety of samples became available; this increased measurement complexity doubtless contributed to the small decrease in measurement performance from 1991 through 1994. The development of graphical analysis and communication techniques starting in 1994 may have some role in the more recent measurement performance improvements.

Changes over Time: Newer Measurands. Figures 4 and 5 display the median measurement performance characteristics for the remaining 11 measurands for all M^2QAP exercises where eight or more participants reported data. The "sudden" onset of data for a given measurand only partially reflects improvements in separation technology;[1] it is more directly linked to the addition of the measurand to the response form included with each set of samples. Many participants apparently will not report values for measurands that are not explicitly requested, while others appear to relish expanding their repertoire. All measurands reported by just one participant are separately reported in the exercise summary report. We speculate that this spurs others to report measurements they are already making and/or to quantitatively analyze peaks they have already qualitatively identified. The number of participants reporting values for new measurands often increases quite quickly.

These newer measurands can be divided into four groups on the basis of common trends in their measurement performance characteristics. (1) Discordance dominates apparent imprecision for γ-tocopherol, total and *trans*-lycopene, and total β-cryptoxanthin. Concordance has improved slowly over time for these measurands while apparent precision has remained fairly constant. This epitomic behavior is consistent with a gradual improvement in chromatographic resolution and, for the lycopenes, better-quality calibration materials and handling procedures.[12] (2) Discordance and apparent imprecision are roughly equal for *cis*-β-carotene and total zeaxanthin, with both measurement characteristics improving slowly over time. These measurands are frequently associated with relatively small peaks eluting on the "tailing edge" of larger peaks. We speculate that the slowly improving measurement performance may result from improved

(12) Stacewicz-Sapuntzakis, M., Pitfalls of Lycopene Measurement, NIST Micronutrients Measurement Quality Assurance Workshop, Experimental Biology '99, April 16, 1999; NIST: Washington, DC, 1999.

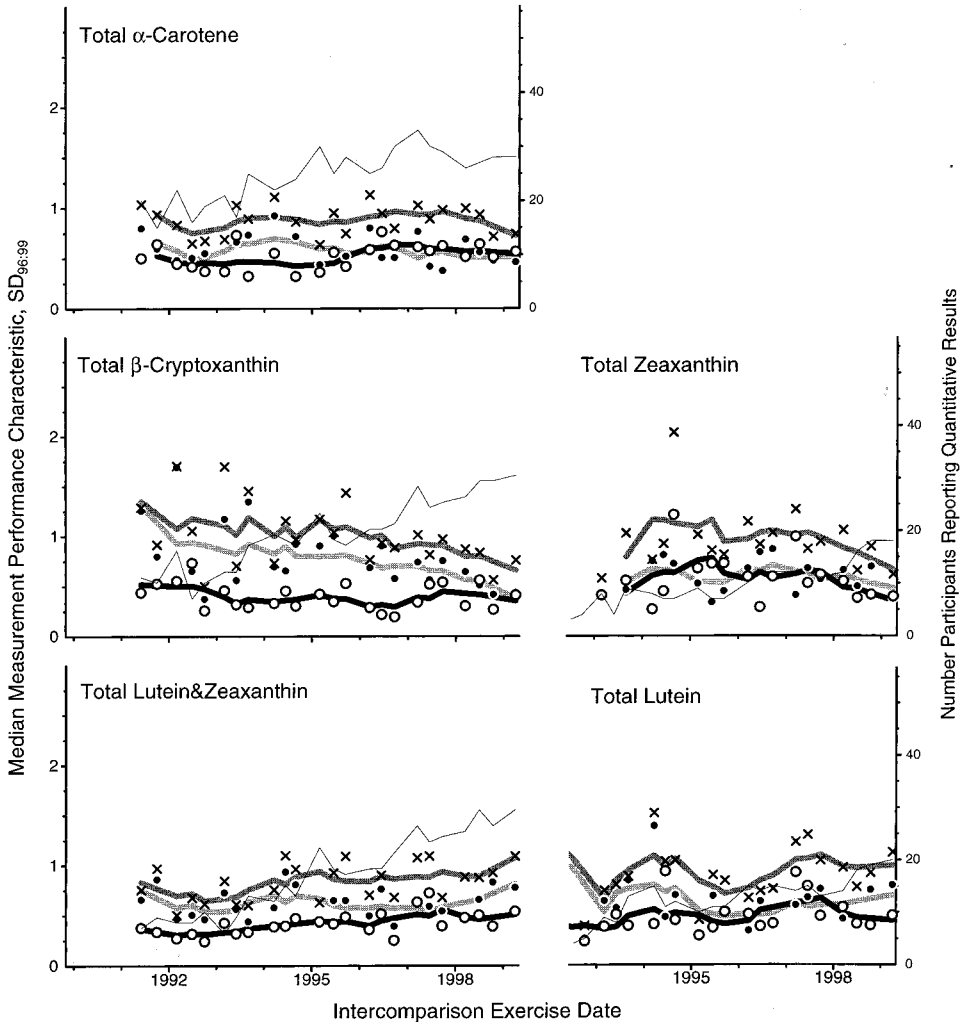

Figure 5. Measurement performance characteristics as a function of time for total α-carotene, total *β*-cryptoxanthin, total lutein, total lutein and zeaxanthin, and total zeaxanthin. Legend as in Figure 3.

chromatographic resolution and/or peak integration protocols. (3) Discordance and apparent imprecision are also roughly equal for *trans-β*-carotene, total α-carotene, and retinyl palmitate, but neither characteristic has changed much over time. These three measurands are determined from peaks that are among the last to elute in typical reversed-phase chromatographic systems. Given that the absolute SD(*X*) for *trans-β*-carotene is consistently better (i.e., smaller) than that for total *β*-carotene,[2] the absence of relative improvement in the *trans-β*-carotene performance characteristics suggests that improving current measurement performance may require use of different measurement methods for these very nonpolar measurands. (4) Both discordance and apparent imprecision have increased over time for total lutein and total lutein and zeaxanthin. Lutein and zeaxanthin isomers typically elute at about the same time as many of the less commonly quantified polar dietary carotenoids and/or metabolites present in serum.[13] We believe that as individual participants improved chromatographic resolution enough to resolve the major (*trans*-) lutein and zeaxanthin peaks, the mix of interferents in these measurands has become more diverse.

Implications. If improved comparability for retinol is desired, the small but consistent dominance of discordance (Figure 4, top) suggests that careful discrimination between "total" and *trans*-retinol measurands will be productive. Since retinol isomers spiked into serum in the laboratory are relatively soluble, it is possible to tailor the relative isomeric composition of samples to enable verification of the measurand identity. If improved comparability for α-tocopherol is desired, the similar if unexpected dominance of discordance over apparent precision (Figure 4, middle) suggests that participants are reporting somewhat different entities as the same measurand. However, the *δ*-, *γ*-, and/or *β*-tocopherol composition of many M[2]QAP samples has been manipulated without apparent effect on reported α-tocopherol values. Addressing α-tocopherol discordance will thus first require identification of the suspected interferences, perhaps starting with the components of common Vitamin E supplements.[14]

In contrast, the consistently equal magnitudes of discordance and apparent precision for total *β*-carotene (Figure 4, bottom) confirms that within-participant quality control is the limiting comparability factor.[2] This general consensus on measurand identity suggests that manipulation of *β*-carotene isomer composi-

(13) Khachik, F.; Spangler, C. J.; Smith, J. C., Jr.; Canfield, L. M.; Steck, A.; Pfander, H. *Anal. Chem.* **1997**, *69*, 1873−81.

(14) Thakur, M. L.; Srivastava, U. S. *Nutr. Res. (N.Y.)* **1996**, *16*, 1767−1809.

tion is unlikely to help improve comparability—even, given their very low serum solubility, were it not difficult to accomplish.

Manipulation of sample composition becomes technically more difficult as measurand solubility in serum decreases; however, for some measurands it is possible to selectively augment samples to enable identification of major sources of discordance. While considerably more expensive than augmentation, a fairly wide range of "challenge samples" can be created by blending selected serum pools natively high in desired measurands or potential interferences. Using such samples to challenge measurement systems (and analysts) does not ensure improved measurement comparability, but identification of the limitations of current measurement systems is a step toward that goal.

"Adequate" comparability is in the eye of the beholder and will be variously defined by participants with different information needs and for measurands associated with compounds of different properties. The expected serum concentration range of *trans*-retinol is quite narrow (due to active physiological regulation through transport proteins), while the expected range for diet-dependent trans-β-carotene is quite large. The relatively concordant "retinol" measurements are thus too discordant to have much epidemiological utility, whereas the relatively discordant β-carotene measurements are more concordant than required for stable comparisons across time and/or among laboratories.[10] Thus, the "most important" problems of interest to the M^{2}QAP community cannot be identified from measurement performance characteristics alone. However, quantitative evaluation of performance characteristics can guide the efficient allocation of resources once critical needs are identified.

ACKNOWLEDGMENT

We thank Dr. Hung-Kung Liu for his attempts to rescue us from the darkest statistical pits we have stumbled into, Dr. William A. MacCrehan for his constructive deconstruction, Dr. Willie E. May for encouraging the evolution of and continuing the financial support of the M^{2}QAP, and the National Cancer Institute for their support during the program's first 12 years. We thank all analysts who have participated in the M^{2}QAP exercises and workshops; their participation, enthusiastic support, and critical feedback are crucial to the continued improvement of measurement comparability.

Received for review December 28, 1999. Accepted June 27, 2000.

AC991480J