Investigating the Current Harmonization Status of Tumor Markers Using Global External Quality Assessment Programs: A Feasibility Study

Huub H. van Rossum,[a,*] Stefan Holdenrieder,[b,c] Bart E.P.B. Ballieux,[d] Tony C. Badrick,[e] Yeo-Min Yun,[f] Chuanbao Zhang,[g] Dina Patel,[h] Marc Thelen,[i,j] Junghan Song,[k] Nathalie Wojtalewicz,[c] Nick Unsworth,[l] Hubert W. Vesper,[m] Wei Cui,[n] Lakshmi V. Ramanathan,[o] Catharine Sturgeon,[l] and Qing H. Meng[p]

BACKGROUND: The harmonization status of most tumor markers (TMs) is unknown. We report a feasibility study performed to determine whether external quality assessment (EQA) programs can be used to obtain insights into the current harmonization status of the tumor markers α-fetoprotein (AFP), prostate specific antigen (PSA), carcinoembryonic antigen (CEA), cancer antigen (CA)125, CA15-3 and CA19-9.

METHODS: EQA sample results provided by 6 EQA providers (INSTAND [Germany], Korean Association of External Quality Assessment Service [KEQAS, South Korea], National Center for Clinical Laboratories [NCCL, China], United Kingdom National External Quality Assessment Service [UK NEQAS, United Kingdom], Stichting Kwaliteitsbewaking Medische Laboratoriumdiagnostiek [SKML, the Netherlands], and the Royal College of Pathologists of Australasia Quality Assurance Programs [RCPAQAP, Australia]) between 2020 and 2021 were used. The consensus means, calculated from the measurement procedures present in all EQA programs (Abbott Alinity, Beckman Coulter DxI, Roche Cobas, and Siemens Atellica), was used as reference values. Per measurement procedure, the relative difference between consensus mean for each EQA sample and the mean of all patient-pool–based EQA samples were calculated and compared to minimum, desirable, and optimal allowable bias criteria based on biological variation.

RESULTS: Between 19040 (CA15-3) and 25398 (PSA) individual results and 56 (PSA) to 76 (AFP) unique EQA samples were included in the final analysis. The mean differences with the consensus mean of patient-pool–based EQA samples for all measurement procedures were within the optimum bias criterion for AFP, the desirable bias for PSA, and the minimum bias criterion for CEA. However, CEA results <8 μg/L exceeded the minimum bias criterion. For CA125, CA15-3, and CA19-9, the harmonization status was outside the minimum bias criterion, with systematic differences identified.

CONCLUSIONS: This study provides relevant information about the current harmonization status of 6 tumor markers. A pilot harmonization investigation for CEA, CA125, CA15-3, and CA19-9 would be desirable.

Introduction

Circulating blood-based tumor markers (TMs) are important diagnostic tools in cancer care. For their optimal clinical use, including application of general clinical decision limits, appropriate harmonization status of the measurement procedures is essential. Three important

[a]Department of Laboratory Medicine, Netherlands Cancer Institute, Amsterdam, the Netherlands; [b]Institute of Laboratory Medicine, Munich Biomarker Research Center, Deutsches Herzzentrum München, Technische Universität München, Munich, Germany; [c]INSTAND e.V., Society for Promoting Quality Assurance in Medical Laboratories, Duesseldorf, Germany; [d]Department of Clinical Chemistry, Leiden University Medical Center, Leiden, the Netherlands; [e]RCPA Quality Assurance Programs, St Leonards, Sydney, Australia; [f]Department of Laboratory Medicine, Konkuk University Medical Center, Seoul, South Korea; [g]National Center for Clinical Laboratories, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing Hospital/National Center of Gerontology, Beijing, China; [h]UK NEQAS Immunology, Immunochemistry & Allergy, Northern General Hospital, Sheffield, United Kingdom; [i]SKML, Nijmegen, the Netherlands; [j]Department of Laboratory Medicine of the Radboud University Medical Center, Nijmegen, the Netherlands; [k]Department of Laboratory Medicine, Seoul National University Bundang Hospital and College of Medicine, Seongnam, South Korea; [l]UK NEQAS [Edinburgh], Department of Laboratory Medicine, Royal Infirmary of Edinburgh, Edinburgh, United Kingdom; [m]Division of Laboratory Sciences, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, GA, United States; [n]Department of Laboratory Medicine, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China; [o]Clinical Chemistry Service, Department of Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, United States; [p]Department of Laboratory Medicine, Division of Pathology and Laboratory Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, United States.

*Address correspondence to this author at: Department of Laboratory Medicine, Netherlands Cancer Institute, Plesmanlaan 121, 1066CX Amsterdam, the Netherlands. Tel 31-20-5122756; E-mail h.v.rossum@nki.nl.

Received September 25, 2023; accepted December 11, 2023.
https://doi.org/10.1093/clinchem/hvae005

tumor markers generally used in cancer care are prostate specific antigen (PSA), carcinoembryonic antigen (CEA), and α-fetoprotein (AFP), for which International Standards (ISs) International Reference Reagent (IRR) 96/670, International Reference Preparation (IRP) 73/601, and IS 72/225, respectively, have been available for many years (1, 2). PSA is essential for the management of prostate cancer, with applications including screening, diagnosis, treatment, and follow-up (3–5). Clinical guidelines for prostate cancer include specific PSA concentration-based decision limits for various clinical applications. CEA is an important tumor marker for advanced colon cancer, breast cancer, and lung cancer. Some recent clinical guidelines for lung and colon cancers do not mention the use of CEA, and guidelines for lung cancer have recommended against its use, primarily based on clinical pathways that do not include today's targeted and immunotherapy-based treatments (6–8). More recent research has clearly demonstrated relevant use of CEA in lung cancer, but current applications for targeted and immunotherapy follow-up have not been appropriately validated (9, 10). Interestingly, the IS available for CEA is not generally used because of concerns regarding its commutability and the units in which CEA results are reported (1, 11). AFP has a critical role in the diagnosis, staging of, and follow-up of hepatocellular carcinoma, and clinical decision concentrations are included in clinical guidelines (12). Three other generally available tumor markers for which no IS is available are cancer antigen (CA)125, CA15-3, and CA19-9. These TMs have clinical applications primarily in ovarian, breast, and gastrointestinal cancers, respectively.

Unfortunately, most TMs are not standardized and their current harmonization status is unknown (1, 13). Harmonization of circulating TMs is challenging for several reasons, including their heterogeneity and the lack of knowledge about which isoform is most clinically relevant, differences in immunoassay design and antibody epitope recognition in the measurement procedures available, and lack of accurate calibration against an appropriate, universal, and commutable international standard (1, 14). The lack of TM harmonization limits interpretation of clinical validation studies, as between measurement procedure differences are often not taken into account. Several TMs lack clinical validation studies that provide a high level of evidence for their clinical use. Such studies are essential to allow any recommendation in evidence-based clinical guidelines. This has led to the removal of TMs from clinical guidelines, such as those for advanced lung cancer and breast cancer (8, 15). In clinical oncology, there is an increasing need for TM measurements e.g., for treatment follow-up and detection of response/nonresponse to treatment. This trend is driven by the increasing availability of new and effective systemic treatments.

To enable comparison of analytical results using different measurement procedures both in clinical research studies and in patient care, appropriate harmonization is essential. Therefore, the purpose of this study was to investigate the feasibility of using global external quality assessment programs (EQA) to investigate the current harmonization status of AFP, CA15-3, CA19-9, CA125, CEA, and PSA.

Materials and Methods

EQA PROVIDERS AND DATA

Data from 6 EQA providers: INSTAND (Germany), Korean Association of External Quality Assessment Service (KEQAS, South Korea), National Center for Clinical Laboratories (NCCL, China), United Kingdom National External Quality Assessment Service (UK NEQAS, United Kingdom), Stichting Kwaliteitsbewaking Medische Laboratoriumdiagnostiek (SKML, the Netherlands), and the Royal College of Pathologists of Australasia Quality Assurance Programs (RCPAQAP, Australia) were included in this study. The PSA, CEA, AFP, CA125, CA15-3, and CA19-9 EQA results for specimens issued during 2020 and 2021 were requested from each EQA provider. Data either included the median (preferred) or mean of TM results for each specified measurement procedure for every individual EQA sample. In addition, the number of laboratories participating with a specified measurement procedure was included for every TM and also a brief description of the characteristics of the EQA materials. Table 1 provides an overview of the characteristics of the included EQA samples for each TM.

STATISTICAL PLAN AND DATA ANALYSIS

To enable comparisons between the EQA programs for each TM, only measurement procedures that were available in all EQA programs were included in the analysis. Since the categorizations and definitions of the measurement procedures in the EQA programs differed, one representative measurement procedure from each manufacturer was selected to reflect current product lines or having the highest number of participating laboratories. Measurement procedures separated in the EQA program that might reasonably be expected to be similar were not merged and only a single measurement procedure as defined by the EQA program was used. A consensus mean was used as reference measure; therefore, the mean of the included individual measurement procedure median (or mean) values, per EQA sample, was calculated. This was done using a non-weighted simple mean calculation to ensure a consistent basis for the consensus mean value and to make sure it is not affected by the relative measurement procedure composition within an EQA program. For the individual measurement procedures, the

Table 1. Tumor marker EQA programs and characteristics of included samples.

EQA program	INSTAND	KEQAS	NCCL	UK NEQAS	RCPAQAP[a]	SKML
Country	Germany	South Korea	China	United Kingdom	Australia	The Netherlands
PSA						
EQA samples (years)	6 (2020–2021)	12 (2020–2021)	10 (2021)	24 (2021) Excl 10[b,c,d]	6 (2021)	12 (2021) Excl 4[c]
EQA material	Spiked serum/plasma	Pooled patient serum	Spiked serum/plasma	Pooled patient serum	Spiked serum/plasma	Pooled patient serum
Results per EQA sample	80	150	1400	163	148	111
CEA						
EQA samples (years)	6 (2020–2021)	12 (2020–2021)	10 (2021)	28 (2021)	6 (2021)	12 (2021)
EQA material	Spiked serum/plasma	Pooled patient serum	Spiked serum/plasma	Pooled patient serum	Spiked serum/plasma	Pooled patient serum
Results per EQA sample	86	156	1589	151	348	67
AFP						
EQA samples (years)	6 (2020–2021)	12 (2020–2021)	10 (2021)	30 (2020)	6 (2021)	12 (2021)
EQA material	Spiked serum/plasma	Pooled patient serum	Spiked serum/plasma	Pooled patient serum	Spiked serum/plasma	Pooled patient serum
Results per EQA sample	49	170	1596	119	144	87
CA125						
EQA samples (years)	6 (2020–2021)	12 (2020–2021)	10 (2021)	20(2021) Excl 2[b]	6 (2021)	12 (2021)
EQA material	Spiked serum/plasma	Commercial iQC	Spiked serum/plasma	Pooled patient serum	Spiked serum/plasma	Pooled patient serum
Results per EQA sample	47	105	1579	55	340	55
CA15-3						
EQA samples (years)	6 (2020–2021)	8 (2020–2021). Excl 4[f]	10 (2021)	20(2021) Excl 2[b]	6 (2021)	12 (2021)
EQA material	Spiked serum/plasma	Commercial iQC	Spiked serum/plasma	Pooled patient serum	Spiked serum/plasma	Pooled patient serum
Results per EQA sample	50	55	1509	97	144	50
CA19-9						
EQA samples (years)	6 (2020–2021)	12 (2020–2021)	10 (2021)	16(2021) Excl 4[b,e]	6 (2021)	12 (2021)
EQA material	Spiked serum/plasma	Commercial iQC	Spiked serum/plasma	Pooled patient serum	Spiked serum/plasma	Pooled patient serum
Results per EQA sample	68	126	1591	147	324	39

Excl, The number of EQA samples excluded from the analysis.

[a] The mean of the same sample distributed and analyzed in quadruplicate was used.

[b] Duplicate or dilution of other sample.

[c] Below 0.1 µg/L.

[d] IS 17/100-based sample.

[e] Very low numerical value (compromises accuracy).

[f] Data of one measurement procedure is not available.

Table 2. Desirable performance specifications.[a]

	Total allowable bias		
	Minimum	Desirable	Optimal
PSA	16.0%	10.6%	5.3%
CEA	22.4%	14.9%	7.5%
AFP	20.8%	13.8%	6.9%
CA125	10.1%	6.7%	3.4%
CA15-3	13.9%	9.3%	4.6%
CA19-9	21.6%	14.4%	7.2%

[a]Performance specifications were based on biological variation data extracted from the EFLM Biological Variation Database and CA19-9 for a recent study performed by an EFLM working group (16, 17).

mean or median value of every EQA sample was then expressed as a percentage (%) of the consensus mean value. These values were plotted and color-coded for each EQA program. In addition, the mean of all individual EQA samples was calculated for every measurement procedure using patient-pool EQA samples only. Finally, the mean difference compared with the consensus mean value was calculated. It was assumed that the EQA samples, based on patient-pool materials, best reflected the behavior of individual patient samples.

The results were interpreted with respect to analytical performance specifications (APS) based on biological variation, as previously described and documented in the European Federation of Laboratory Medicine (EFLM) Biological Variation database (16, 17), using the minimum, desirable and optimal specifications for bias. An overview of the criteria used, per TM, is presented in Table 2.

Results

SELECTION OF MEASUREMENT PROCEDURES AND DEFINITION OF THE CONSENSUS MEAN

As results measurement procedures for the Abbott (Alinity), Beckman (DxI), Roche (Cobas), and Siemens (Atellica) were available from all included EQA programs, these measurement procedures were selected and used to calculate the consensus mean to enable comparisons between EQA programs. For some TMs, only the abovementioned manufacturer names were available in EQA program results. In these cases, measurement procedures were referred to as Abbott Alinity/Architect, Beckman Access/DxI, Roche Cobas, and Siemens Atellica/Centaur.

For each TM, from 19040 (CA15-3) to 25398 (PSA) individual laboratory results were included and used in the final analyses. Three EQA programs (INSTAND, NCCL, and RCPAQAP) used serum/plasma spiked with exogenous materials to obtain sufficient EQA volumes and relevant TM concentrations. Two EQA programs (SKML and UK NEQAS) only used pooled patient sera to obtain elevated TM concentration, and one EQA program used both pooled patient sera as well as commercial internal quality control (iQC) materials depending on the tumor marker (KEQAS). An overview of the EQA program and sample characteristics is presented in Table 1. The number of EQA samples (including patient-pool–based EQA samples) included per tumor marker were: PSA 56 (34), CEA 74 (52), AFP 76 (54), CA125 64 (30), CA15-3 60 (30), and CA19-9 58 (24). An overview of the obtained measurement procedure mean expressed as percentage of the consensus mean for each individual EQA program is presented in Table 3.

PSA

Results of the PSA harmonization study are presented in Fig. 1 for the measurement procedures included in the consensus mean calculation. For Beckman Coulter, the calibration (WHO or Hybritech) was only specified by the UK NEQAS scheme, which used the WHO calibration. Only 5 EQA samples from the UK NEQAS program were outside the minimum bias criterion. The average patient-pool–based EQA difference with the consensus mean was −0.3% for Abbott Alinity, −2.7% for Beckman Coulter DxI, 10.0% for Roche Cobas, and −7.9% for Siemens Atellica, all within the desirable allowable bias criterion. Furthermore, for Siemens Atellica a trend towards a negative bias at the low concentration range was observed. The spiked serum/plasma-based EQA samples and patient-pool–based EQA samples appeared to behave rather similarly.

CEA

The CEA harmonization investigation is presented in Fig. 2 for measurement procedures included in the consensus mean calculation. The average difference with the consensus mean was 7.4% for Abbott Alinity, 8.2% for Beckman Access/DxI, 3.7% for Roche Cobas, and −19.4% for Siemens Atellica. Although all these mean recoveries were within the minimum allowable bias criterion of ±22.4%, several individual EQA results were outside this criterion. Particularly at the concentration range below 8 µg/L, all patient-pool–based EQA samples showed a trend towards an increasing negative difference, with the consensus mean exceeding the minimum bias criterion for the Siemens Atellica method.

AFP

The AFP harmonization investigation is presented in Fig. 3. The average patient-pool EQA difference from

Table 3. EQA program specific TM harmonization investigation result.[a]

	INSTAND	KEQAS	NCCL	UK NEQAS	RCPAQAP	SKML	Mean patient pool EQA
PSA							
Abbott	98.0 (95.1–100.9)	**95.5 (92.6–98.3)**	92.3 (89.8–94.8)	**104.6 (101.5–107.6)**	92.3 (88.9–95.7)	**97.6 (92.9–102.4)**	**99.7 (97.4–102.1)**
Beckman	103.4 (100.2–106.6)	**102.7 (100.7–104.8)**	103.2 (102.1–104.2)	**89.1 (85.3–93.0)**	109.2 (106.7–111.6)	**103.5 (97.5–109.6)**	**97.3 (94.1–100.5)**
Roche	105.3 (102.7–107.8)	**111.1 (109.8–112.4)**	105.6 (105.3–105.9)	**110.1 (109.2–111.1)**	103.0 (100.0–106.0)	**108.2 (107.2–109.1)**	**110.0 (109.3–110.7)**
Siemens	93.3 (90.9–95.8)	**90.7 (89.0–92.3)**	98.9 (97.2–100.7)	**96.2 (93.9–98.5)**	95.5 (93.9–97.1)	**90.7 (88.4–92.9)**	**92.9 (91.4–94.4)**
CEA							
Abbott	119.3 (114.2–124.5)	**108.7 (105.2–112.3)**	120.2 (115.3–125.1)	**106.2 (100.4–112.0)**	125.0 (120.0–130.1)	**109.1 (107.7–110.4)**	**107.4 (104.2–110.7)**
Beckman	97.9 (96.5–99.3)	**109.8 (105.9–113.6)**	92.9 (92.0–93.7)	**109.6 (106.2–113.1)**	90.9 (88.5–93.3)	**103.5 (101.8–105.2)**	**108.2 (106.0–110.5)**
Roche	82.1 (78.6–85.6)	**102.2 (99.0_105.4)**	84.6 (81.1–88.1)	**103.3 (97.4–109.2)**	90.1 (82.7–97.4)	**106.1 (102.2–110.1)**	**103.7 (100.3–107.1)**
Siemens	100.6 (93.6–107.6)	**79.3 (76.7–82.0)**	102.3 (99.4–105.3)	**80.9 (76.1–85.7)**	94.0 (91.9–96.0)	**81.3 (77.0–85.6)**	**80.6 (77.8–83.4)**
AFP							
Abbott	98.5 (95.8–101.2)	**97.5 (96.7–98.3)**	94.3 (93.4–95.3)	**96.6 (96.0–97.2)**	96.4 (90.5–102.3)	**96.1 (94.6–97.5)**	**96.7 (96.2–97.2)**
Beckman	95.8 (92.0–99.7)	**93.9 (93.2–94.1)**	93.3 (92.5–94.1)	**95.6 (94.4–96.9)**	100.5 (97.4–103.7)	**97.2 (93.9–100.5)**	**95.6 (94.6–96.6)**
Roche	95.0 (91.2–98.8)	**104.8 (103.1–105.7)**	104.4 (103.1–105.7)	**103.8 (102.7–104.9)**	96.7 (94.6–98.9)	**102.3 (100.3–104.2)**	**103.7 (102.9–104.5)**
Siemens	110.7 (104.6–116.7)	**103.8 (102.4–105.2)**	108.0 (106.4–109.5)	**104.1 (103.1–105.0)**	106.4 (105.3–107.5)	**104.5 (102.7–106.2)**	**104.1 (103.4–104.8)**
CA125							
Abbott	129.6 (123.4–135.8)	105.9 (104.1–107.6)	122.5 (117.1–127.9)	**121.5 (119.2–123.7)**	122.8 (119.8–125.8)	**115.9 (113.5–118.4)**	**119.3 (117.4–121.2)**
Beckman	73.7 (63.8–83.7)	130.8 (126.4–135.2)	43.8 (42.2–45.4)	**81.2 (76.5–85.8)**	68.7 (63.6–73.7)	**103.4 (100.1–106.6)**	**90.1 (85.1–95.0)**
Roche	76.2 (72.8–79.6)	70.9 (68.4–73.4)	130.1 (126.4–133.7)	**87.8 (81.6–94.0)**	89.4 (83.5–95.3)	**91.9 (87.0–96.8)**	**89.4 (85.2–93.6)**
Siemens	120.4 (117.5–123.3)	92.5 (88.0–97.0)	103.7 (101.7–105.7)	**109.5 (105.9–113.1)**	119.2 (116.5–121.9)	**88.8 (87.5–90.0)**	**101.2 (96.9–105.5)**
CA 15–3							
Abbott	108.6 (103.3–113.8)	107.9 (104.7–111.1)	137.7 (130.3–145.2)	**114.8 (108.9–120.7)**	152.3 (136.5–168.1)	**102.6 (101.1–104.2)**	**109.8 (105.6–113.9)**
Beckman	58.9 (53.3–64.5)	59.7 (54.3–65.0)	39.1 (34.6–43.6)	**68.2 (62.9–73.5)**	49.7 (41.1–58.3)	**74.3 (71.6–76.9)**	**70.7 (67.3–74.2)**
Roche	107.6 (104.4–110.8)	116.4 (115.2–117.6)	113.1 (109.8–116.3)	**107.2 (104.1–110.4)**	70.0 (54.0–86.0)	**113.2 (112.4–113.9)**	**109.7 (107.5–111.8)**
Siemens	124.9 (120.0–129.8)	116.0 (111.6–120.5)	110.1 (108.8–111.4)	**109.8 (106.3–113.3)**	128.1 (117.0–139.1)	**109.9 (108.0–111.9)**	**109.9 (107.7–112.0)**
CA 19–9							
Abbott	206.7 (185.5–227.9)	100.8 (96.8–104.8)	76.6 (74.3–78.9)	**204.8 (190.6–218.9)**	248.8 (244.9–252.8)	**150.4 (112.9–187.9)**	**177.6 (155.1–200.1)**
Beckman	62.0 (53.5–70.6)	109.5 (106.7–112.3)	191.4 (179.7–203.1)	**67.4 (64.7–70.2)**	40.6 (39.1–42.1)	**92.2 (86.9–97.5)**	**79.8 (74.0–85.7)**
Roche	40.2 (33.6–46.7)	88.7 (86.6–90.8)	75.7 (71.3–80.1)	**38.8 (33.3–43.7)**	38.3 (37.3–39.3)	**65.0 (42.7–87.2)**	**51.9 (39.5–64.2)**
Siemens	91.1 (83.7–98.6)	101.0 (96.7–105.2)	56.3 (48.5–64.2)	**89.0 (81.9–96.2)**	72.3 (67.4–77.1)	**92.4 (80.9–103.9)**	**90.7 (84.1–97.3)**

[a] The measurement procedure mean of all EQA samples within an EQA program is expressed as % of the consensus mean. Bold marked fields are included in the mean patient pool EQA calculations. Values within parentheses represent the 95% CI of the calculated mean value.

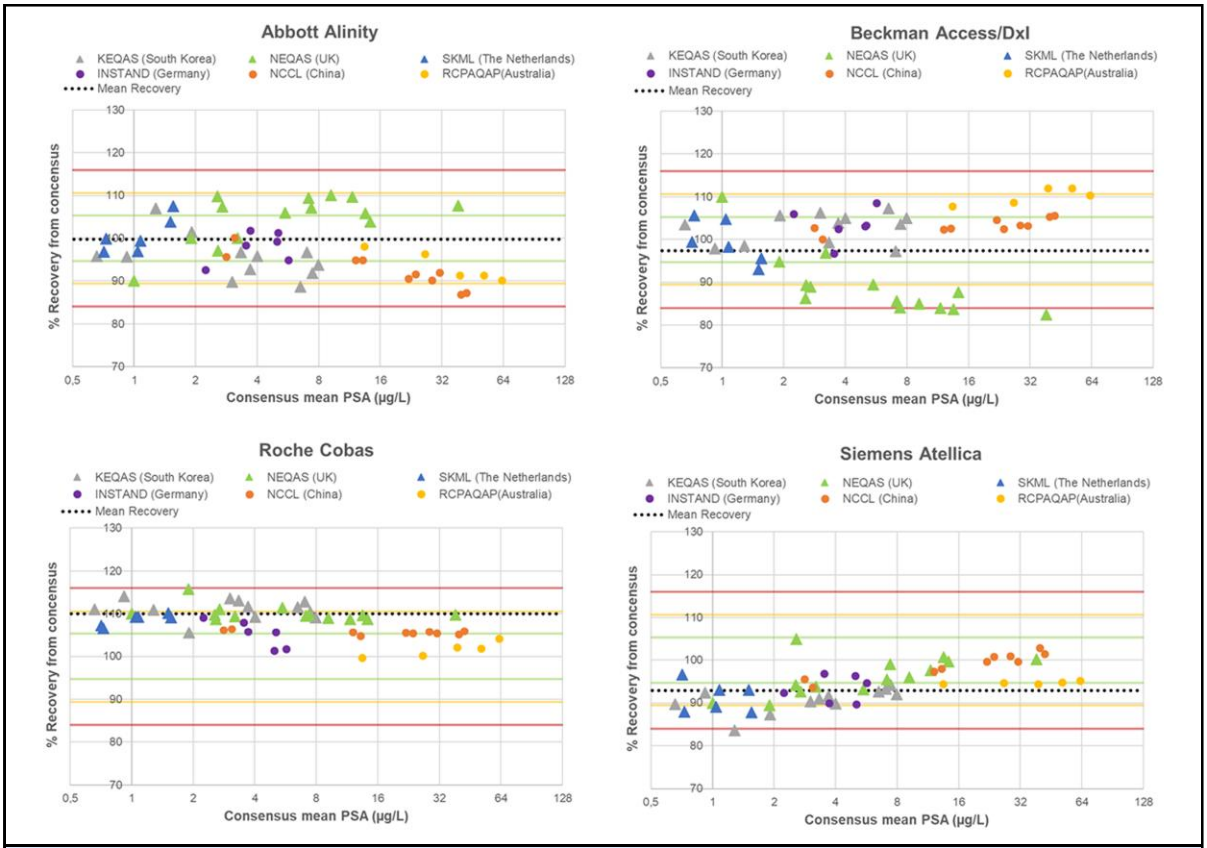

Fig. 1. PSA harmonization investigation. Patient-pool–based EQA samples are plotted as triangles and spiked samples as circles. Mean recovery is calculated using patient-pool–based EQA samples only. Green, yellow, and red lines represent the optimum, desirable, and minimum allowable bias criteria, respectively.

the consensus mean was −3.3% for Abbott Alinity, −4.4% for Beckman DxI, 3.7% for Roche Cobas, and 4.1% for Siemens Atellica, all within the optimal bias criterion of ±6.9%. All EQA samples were within the minimum bias criterion of ±20.8% and all but 3 were within the desirable bias criterion of ±13.8%.

CA125, CA15-3, AND CA19-9

The results for CA125, CA15-3, and CA19-9 are presented in online Supplemental data 1–3. For each of these TMs, the samples from the individual EQA programs seem to behave rather differently and often provided discrepant and contrary results between measurement procedures. When focusing on the patient-pool–based EQA, this suggested that for CA15-3 the Beckman Access/DxI measurement procedures provide significantly lower concentrations than the consensus values with a mean difference of −29% from the consensus mean, while the other 3 had a similar mean differences from the consensus mean of approximately 10%.

For CA125, the Abbott Alinity/Architect measurement procedures seemed to have a positive bias with respect to the consensus mean of 19%. For the other 3 CA125 measurement procedures, individual patient-pool EQA results had differences higher and lower than the minimum allowable bias criterion.

For CA19-9, the Abbott Alinity/Architect measurement procedures produced significantly higher results with a mean difference of 77% from the consensus mean, while Roche Cobas seemed to produce significantly lower results with a mean difference of −48% when compared to the consensus mean.

For the investigated measurement procedures of CA 15-3, CA125, and CA19-9, the minimum bias criterion of ±13.9%, ± 10.1%, and ±21.6% respectively, were exceeded.

Discussion

This study investigated the feasibility of using several global EQA programs to examine the harmonization

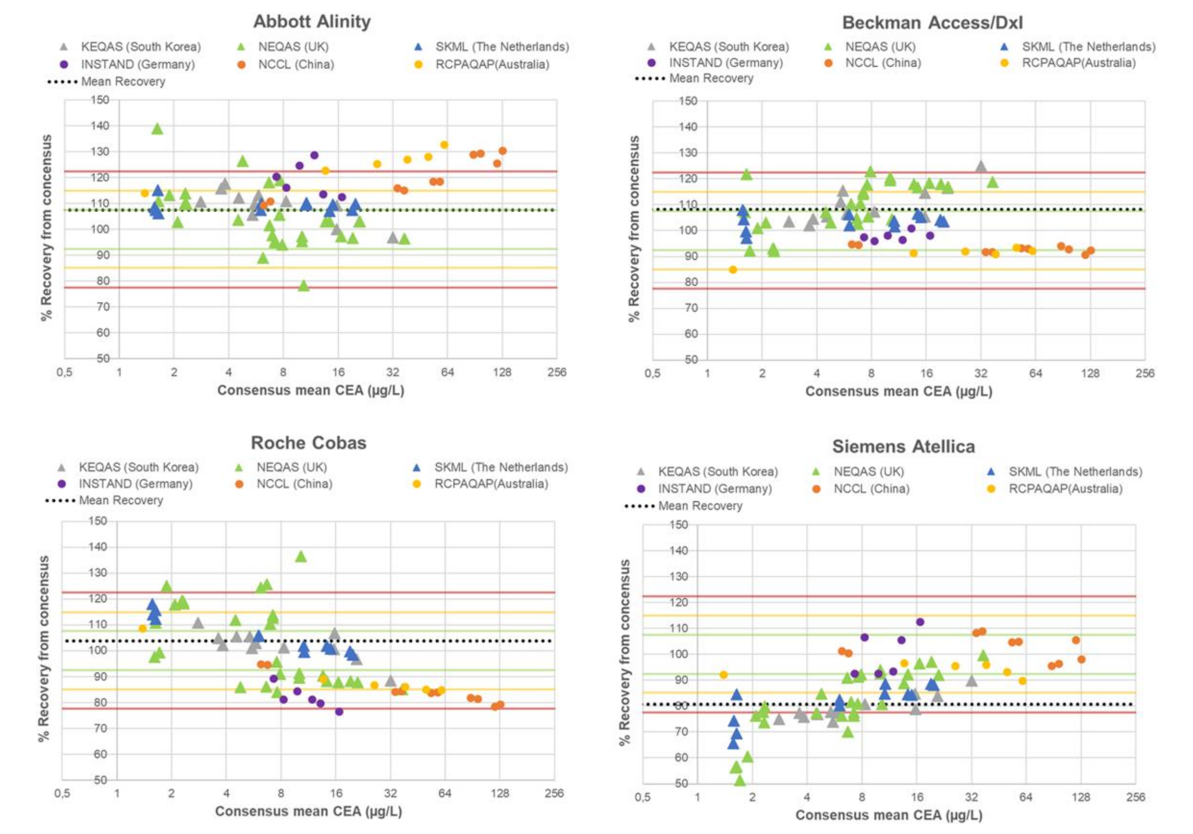

Fig. 2. CEA harmonization investigation. Patient-pool–based EQA samples are plotted as triangles and spiked samples as circles. Mean recovery is calculated using patient-pool–based EQA samples only. Green, yellow, and red lines represent the optimum, desirable, and minimum allowable bias criteria, respectively.

status of 6 widely used TMs. Results of this study may help to prioritize the need to harmonize the different TMs. Based on the patient-pool–based EQA samples, AFP seems to be harmonized within the optimal bias criterion (±6.9%), PSA within the desirable bias criterion (±10.6%), and CEA within the minimum bias criterion (22.4%). The current harmonization status of CA125, CA15-3, and CA19-9 is outside the minimum bias criteria of ±10.1%, ± 13.9%, and ±21.6%, respectively.

Investigating the harmonization of in vitro diagnostics using data from EQA programs has recently gained interest as a tool to provide insights into between measurement procedure relationships and correlations (18–21). A major advantage of using EQA data is that, in general, a large number of measurements are performed per measurement procedure and the median (or mean) of each EQA sample thereby reflects true operational performance. A key and essential requirement is that the EQA materials are commutable (19, 22). This is likely to vary and may not have been formally demonstrated by all EQA programs, including those for TMs. Results of the present study suggest that lack of commutability may be particularly problematic for CA125 and CA19-9. Unless commutability of the EQA samples can be demonstrated, the discrepant EQA data should, therefore, not be blindly used for assessment of between measurement procedure agreement as the results will be highly dependent on the EQA program used (23, 24). Having results from multiple EQA programs, including those that use residual patient material as included in the present study, allows for a more complete evaluation. It is relevant for the laboratory community to have this data available and to use it to gain insights into the effect of using EQA materials with limited or, at best, undocumented commutability. In this study, however, we have only included data from the patient-pool–based EQA programs when assessing harmonization status as these samples were thought the most likely to share the characteristics of individual patient samples.

Data from the individual EQA programs based on patient-pool samples suggest that, for CEA, AFP,

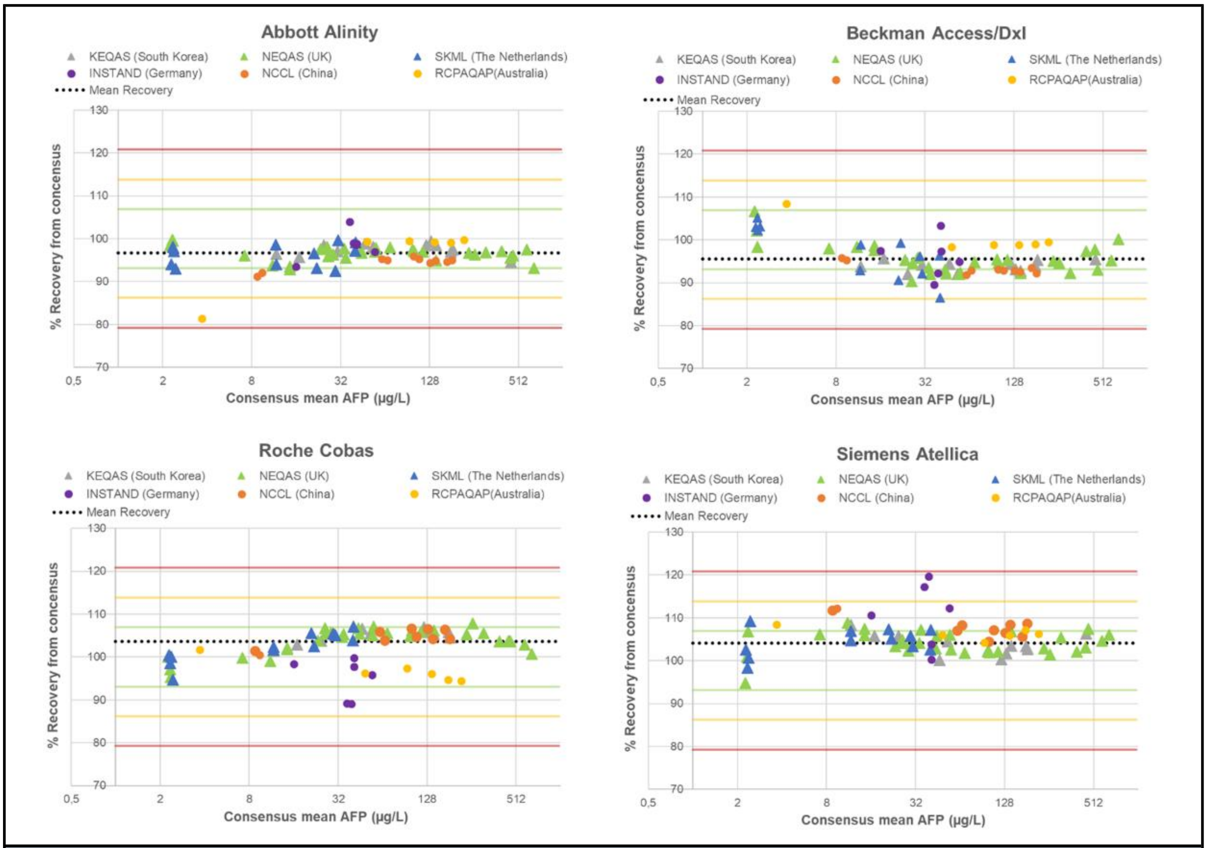

Fig. 3. AFP harmonization investigation. Patient-pool–based EQA samples are plotted as triangles and spiked samples as circles. Mean recovery is calculated using patient-pool–based EQA samples only. Green, yellow, and red lines represent the optimum, desirable, and minimum allowable bias criteria, respectively.

CA15-3, and CA19-9, the EQA samples from the different EQA programs seem to behave rather similarly. For CA125, however, individual EQA samples within or between EQA programs seem to behave differently—these EQA samples gave results which were either higher or lower than the upper and lower criterion for the minimum bias criterion, respectively (online Supplemental 1). These results suggest that spiked or modified materials and patient pools might not provide an adequate EQA material in this case and, preferably, individual patient samples should be used. Alternatively, this might also indicate that CA125 harmonization for the investigated measurement procedures is compromised by the heterogeneity of CA125 in patients, in combination with the differences in immunoassay design. Similarly, for CA19-9, the low SKML EQA samples seem to have different characteristics to the other patient-pool–based EQA samples, also indicating and illustrating differences in immunoassay design and antibody epitope recognition locations of the measurement procedures and potential differences between cancer-derived CA 19-9 and low-level CA 19-9 in healthy persons. However, despite these limitations, the current harmonization for CA125, CA15-3, and CA 19-9 still seems to be outside the minimum allowable bias criterion.

When investigating TM harmonization status using the patient-pool–based EQA only, our results indicate and confirm an adequate harmonization status within the optimum bias criterion (±6.9%) for AFP for the included measurement procedures (13, 25).

For PSA, our results can be compared to a recent harmonization verification study performed by Ferraro et al. for exactly the same measurement procedures and a slightly different consensus value (median instead of mean) (14). Their conclusion was that harmonization amongst PSA measurement procedures was within the minimum allowable bias criterion (±16.0%), but not within the desirable bias criterion (±10.6%). When comparing the mean bias per measurement procedure, the biases for Roche Cobas (11.3% vs 10%) and Siemens Atellica (−7.1% vs −7.1%) from Ferraro et al. and our study, respectively, are highly comparable. However, these are different for Abbott Alinity (6.3% vs −0.3%) and Beckman DxI (−10.3% and −2.7%) as the 95% CIs provided do not overlap. A relevant complicating factor here is the unknown calibration basis of the Beckman DxI measurement procedures included. Others have already demonstrated lower PSA results with the WHO-calibrated Beckman-Coulter Access-II assay in comparison to Roche (Cobas) and Siemens (Centaur) (26). The Beckman DxI EQA results included were probably a mix of both calibrations (WHO and Hybritech), thereby complicating the analysis, but this situation does reflect clinical practice and the true operational harmonization status. Another reason for the differences in harmonization results compared with Ferraro et al. could be the calibration bias when a verification is performed in a single or limited number of analytical runs and reagent lots (14). The use of multiple EQA programs and laboratories within EQA programs would average out any individual calibration bias. This is a major advantage of using EQA programs for method comparison studies.

For CEA, the mean recoveries of the patient-based EQA for Abbott Alinity and Roche Cobas were within the minimum (±22.4%) and those for Beckman Access/DxI within the desirable (±11.9%) bias criterion. However, a scatter of the EQA samples was observed. For Siemens Atellica at lower CEA concentrations (<8 µg/L), a negative bias with respect to the consensus mean was observed. Others have investigated the harmonization status of CEA and Zhang et al. found discrepant results when performing method comparison studies based on patient populations, IRP 73/601, and EQA samples indicating non-commutability of the non-patient derived materials (11). Park et al. have performed a thorough method comparison study of similar measurement procedures by the same manufacturers included in our study (27). Although both the included patient-based method comparison studies were analyzed by between measurement procedure comparisons, both showed that on average (assessing the slope from regression analysis) the Siemens Atellica measurement procedure provided the lowest results and Abbott Alinity the highest CEA results. This is in line with our observation that the average recovery from Siemens Alinity was lowest with an average systematic difference of −19.7%. Since, in our analysis, EQA samples including patient-pool samples exceeded the minimum desirable bias criterion (±22.4%), this together with the 2 method comparison studies provides a strong indication that the included CEA measurement procedures are insufficiently harmonized throughout the measurement range.

For CA125, CA15-3, and CA19-9, the patient-pool EQA indicated a harmonization status exceeding the minimum bias criterion. The next step would be to initiate a harmonization pilot study for CEA, CA125, CA15-3, and CA19-9 TMs, ideally based on individual patient samples. However, based on our results the use of patient-pool–based EQA samples might also be a potential approach. Recently, new procedures for harmonization and standardization were published by the International Organization for Standardization (ISO) and the International Federation of Clinical Chemistry and Laboratory medicine (IFCC) (28). ISO 21151:2020 IS, designed to enable harmonization for measurands when no fit-for-purpose certified reference materials or reference measurement procedures are available, might provide the necessary protocol and methodology. Such a harmonization procedure would require several essential steps including demonstration of the commutability of the materials used, appropriate calibration procedures, and result validation using an independent validation cohort.

Several limitations of our study need to be considered. First, as mentioned previously, non-commutability is the most likely explanation for the fact that non-human EQA materials showed different results to human-sample–based EQA materials for some TMs. This does not prove these samples non-commutable, neither does it prove the commutability of the human-based samples. However, based on their comparable between method behavior, the commutability of the latter is more likely. Another limitation is that only measurement procedures of 4 in vitro diagnostic companies were included in the analysis for all TMs, while more measurement procedures from other in vitro diagnostic companies are available. The 4 measurement procedures included for each TM were the only ones available in all participating EQA programs. By including these in the consensus value for all EQA samples, between measurement procedure and between EQA program comparisons were possible; otherwise, the consensus value would not be applicable. In addition, the categorization of the measurement procedures in the different EQA programs was different and could affect the results. For example, some smaller EQA programs only listed the manufacturer name, while others separated the many individual measurement procedures of one supplier (for example, separating measurement procedures for Roche E411, E601, and E801 systems). In the latter case, one representative procedure, based on the largest number of participants, was selected. Siemens, in particular, is known to have measurement procedures of different origin (Dimension series, Centaur series, and Immulite series) that can have significant differences in assay design for the same TM. In addition, actual differences between measurement procedures may have exceeded the observed mean recovery values and minimum bias criterion (e.g., for the measurement procedures with the highest and lowest observed mean recovery percentages, such as for the Roche vs Siemens PSA results). Finally, the relevance of the APS criteria used to determine the harmonization status can be questioned in terms of clinical relevance (29). Although they are based on a methodology commonly used in the field of laboratory medicine and are evidence-based, the minimum bias criteria for CA125 and CA 15-3 in particular (10.1% and 13.9%) seem rather stringent (16, 17, 29).

In conclusion, although true commutability of the materials used was not demonstrated, this study provided relevant insights into the actual harmonization status of PSA, AFP, CEA, CA125, CA15-3, and CA19-9. Our results suggest that AFP is harmonized sufficiently within the optimal bias criterion and that PSA harmonization status is, on average, within the desirable bias criterion. The average CEA harmonization status is within the minimum bias criterion; however, at the lower concentration range (<8 µg/L) CEA harmonization status is outside the minimum bias criterion. We recommend a follow-up study that investigates the possibility of harmonizing CEA, CA125, CA15-3, and CA19-9 according to ISO recommendations.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

Nonstandard Abbreviations: TM, tumor marker; EQA, external quality assessment; AFP, α-fetoprotein; PSA, prostate-specific antigen; CEA, carcinoembryonic antigen; CA, cancer antigen; UK NEQAS, United Kingdom National External Quality Assessment Service; IS, International Standard.

Author Contributions: *The corresponding author takes full responsibility that all authors on this publication have met the following required criteria of eligibility for authorship: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. Nobody who qualifies for authorship has been omitted from the list.*

Huub van Rossum (Conceptualization-Equal, Data curation-Lead, Formal analysis-Lead, Investigation-Lead, Methodology-Lead, Project administration-Equal, Validation-Lead, Writing—original draft-Lead), Stefan Holdenrieder (Methodology-Equal, Writing—review & editing-Equal), Bart E.P.B. Ballieux (Writing—review & editing-Equal), Tony Badrick (Writing—review & editing-Equal), Yeo Min Yun (Writing—review & editing-Equal), chuanbao zhang (Writing—review & editing-Equal), Dina Patel (Writing—review & editing-Equal), Marc Thelen (Writing—review & editing-Equal), Junghan Song (Writing—review & editing-Equal), Nathalie Wojtalewicz (Writing—review & editing-Equal), Nick Unsworth (Writing—review & editing-Equal), Hubert Vesper (Writing—review & editing-Equal), Wei Cui (Writing—review & editing-Equal), Lakshmi Ramanathan (Writing—review & editing-Equal), Catharine Sturgeon (Conceptualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), and Qing Meng (Conceptualization-Equal, Writing—review & editing-Equal).

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form.*

Research Funding: H. van Rossum, Roche, Health Holland. S. Holdenrieder, EU, Roche, Volition, Sysmex. L. Ramanathan, Abbott, Nova Biomedical.

Disclosures: H. van Rossum, support from Roche Diagnostics, Abbott, Beckman, Siemens, Sysmex, Huvaros; employment or leadership roles: Huvaros, Director Huvaros BV, Chief Scientific Officer SelfSafeSure Blood Collections B.V.; honoraria: Huvaros B.V.; expert testimony: Abbott, Siemens; patents: PCT/NL2016/050315 Method for setting measuring equipment, computer program and measuring equipment, NL2019/079641 Blood collection device and method for the self-collection of blood by a user. S. Holdenrieder, employment or leadership roles: Board ISOBM, WHO-IARC Steering Committee, Associate Editor for several oncological and lab journals, SFZ BioCoDE, CEBIO; consultant or advisory roles: Instand, Roche, Thermo, Merck. L. Ramanathan, support from ADLM (formerly AACC) for travel to attend 2022 Annual Scientific Meeting, support from Abbott Diagnostics, 2021–2023 for validating infectious disease markers on the i 2000, support from Nova Biomedical, 2023, to study ionized Mg in critically ill patients.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

References

1. Sturgeon C. Standardization of tumor markers—priorities identified through external quality assessment. Scand J Clin Lab Invest Suppl 2016;245:S94–9.
2. WHO. WHO International Biological Reference Preparations. 2018. https://cdn.who.int/media/docs/default-source/biologicals/blood-products/catalogue/alphabetical-list.pdf?sfvrsn=15455482_2 (Accessed December 2023).
3. Parker C, Castro E, Fizazi K, Heidenreich A, Ost P, Procopio G, et al. Prostate cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2020;31:1119–34.
4. Ferraro S, Bussetti M, Panteghini M. Serum prostate-specific antigen testing for early detection of prostate cancer: managing the gap between clinical and laboratory practice. Clin Chem 2021;67:602–9.
5. Kanesvaran R, Castro E, Wong A, Fizazi K, Chua MLK, Zhu Y, et al. Pan-Asian adapted ESMO Clinical Practice Guidelines for the diagnosis, treatment and follow-up of patients with prostate cancer. ESMO Open 2022;7:100518.
6. Yoshino T, Arnold D, Taniguchi H, Pentheroudakis G, Yamazaki K, Xu R-H, et al. Pan-Asian adapted ESMO consensus guidelines for the management of patients with metastatic colorectal cancer: a JSMO-ESMO initiative endorsed by CSCO, KACO, MOS, SSO and TOS. Ann Oncol 2018;29:44–70.
7. Grunnet M, Sorensen JB. Carcinoembryonic antigen (CEA) as tumor marker in lung cancer. Lung Cancer 2012;76:138–43.
8. Gennari A, André F, Barrios CH, Cortés J, de Azambuja E, DeMichele A, et al. ESMO Clinical Practice Guideline for the diagnosis, staging and treatment of patients with metastatic breast cancer. Ann Oncol 2021;32:1475–95.
9. Holdenrieder S, Wehnl B, Hettwer K, SimonK, Uhlig S, Dayyani F, et al. Carcinoembryonic antigen and cytokeratin-19 fragments for assessment of therapy response in non-small cell lung cancer: a systematic review and meta-analysis. Br J Cancer 2017;116:1037–45.
10. Muller M, Hoogendoorn R, Moritz RJG, van der Noort V, Lanfermeijer M, Korse CM, et al. Validation of a clinical blood-based

decision aid to guide immunotherapy treatment in patients with non-small cell lung cancer. Tumour Biol 2021;43:115–27.

11. Zhang K, Huo H, Lin G, Yue Y, Wang Q, Li J. A long way to go for the harmonization of four immunoassays for carcinoembryonic antigen. Clin Chim Acta 2016;454:15–9.

12. Vogel A, Cervantes A, Chau I, Daniele B, Llovet JM, Meyer T, et al. Hepatocellular carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2018;29:iv238–55.

13. ICHCLR. Status of harmonisation or standardization of measurands. 2022. www.harmonization.net/measurands (Accessed December 2023).

14. Ferraro S, Bussetti M, Rizzardi S, Braga F, Panteghini M. Verification of harmonization of serum total and free prostate-specific antigen (PSA) measurements and implications for medical decisions. Clin Chem 2021;67:543–53.

15. Wu YL, Planchard D, Lu S, Sun H, Yamamoto N, Kim D-W, et al. Pan-Asian adapted clinical practice guidelines for the management of patients with metastatic non-small-cell lung cancer: a CSCO-ESMO initiative endorsed by JSMO, KSMO, MOS, SSO and TOS. Ann Oncol 2019;30:171–210.

16. EFLM. EFLM biological variation database. https://biologicalvariation.eu (Accessed March 2023).

17. Coşkun A, Aarsand AK, Sandberg S, Guerra E, Locatelli M, Díaz-Garzón J, et al. Within- and between-subject biological variation data for tumor markers based on the European Biological Variation Study. Clin Chem Lab Med 2022;60:543–52.

18. van der Hagen EAE, Weykamp C, Sandberg S, Stavelin AV, MacKenzie F, Miller WG. Feasibility for aggregation of commutable external quality assessment results to evaluate metrological traceability and agreement among results. Clin Chem Lab Med 2020;59:117–25.

19. Badrick T, Miller WG, Panteghini M, Delatour V, Berghall H, MacKenzie F, Jones G. Interpreting EQA-understanding why commutability of materials matters. Clin Chem 2022;68:494–500.

20. Badrick T, Punyalack W, Graham P. Commutability and traceability in EQA programs. Clin Biochem 2018;56:102–4.

21. Wojtalewicz N, Vierbaum L, Kaufmann A, Schellenberg I, Holdenrieder S. Longitudinal evaluation of AFP and CEA external proficiency testing reveals need for method harmonization. Diagnostics (Basel) 2023;13:2019.

22. Braga F, Panteghini M. Commutability of reference and control materials: an essential factor for assuring the quality of measurements in laboratory medicine. Clin Chem Lab Med 2019;57:967–73.

23. Miller WG, Schimmel H, Rej R, Greenberg N, Ceriotti F, Burns C, et al. IFCC Working Group recommendations for assessing commutability part 1: general experimental design. Clin Chem 2018;64:447–54.

24. Nilsson G, Budd JR, Greenberg N, Delatour V, Rej R, Panteghini M, et al. IFCC Working Group recommendations for assessing commutability part 2: using the difference in bias between a reference material and clinical samples. Clin Chem 2018;64: 455–64.

25. Partridge K, Moore M, Atkinson E, Rigsby P, Cowper B. The Second WHO International Standard for alpha-fetoprotein (human, native). WHO/BS/2023.2461. Geneva (Switzerland): WHO; 2023.

26. Boegemann M, Arsov C, Hadaschik B, Herkommer K, Imkamp F, Nofer J-R, et al. Discordant prostate specific antigen test results despite WHO assay standardization. Int J Biol Markers 2018;33:275–82.

27. Park J, Lee S, Kim Y, Choi A, Lee H, Lim J, et al. Comparison of four automated carcinoembryonic antigen immunoassays: ADVIA Centaur XP, ARCHITECT I2000sr, Elecsys E170, and Unicel Dxi800. Ann Lab Med 2018;38:355–61.

28. Miller WG, Greenberg N. Harmonization and standardization: where are we now? J Appl Lab Med 2021;6:510–21.

29. van Rossum HH, Meng QH, Ramanathan LV, Holdenrieder S. A word of caution on using tumor biomarker reference change values to guide medical decisions and the need for alternatives. Clin Chem Lab Med 2022;60:553–5.