Quality specifications and their daily application to evaluate the accuracy of reference measurements for serum concentrations of 25-hydroxyvitamin D_3 and 25-hydroxyvitamin D_2

Ekaterina M. Mineva, **Maya R. Sternberg**, **Christine M. Pfeiffer**, **Shahzad S. Momin**, **Khin L. Maw**, **Rosemary L. Schleicher**[*]

Division of Laboratory Sciences, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, GA, 30341

Abstract

Background: Reference measurement procedures (RMP) have rigorous accuracy specifications. For total 25-hydroxyvitamin D, 25(OH)D, bias 1.7% and CV 5% are recommended. These quality specifications are impractical for minor analytes, such as 25(OH)D_2. Furthermore, documentation on RMP quality performance specifications for the individual 25(OH)D metabolites and their daily application are missing.

Methods: To assess accuracy, we used zeta-scores. Daily, 5–10 specimens (duplicate) and 3 reference materials (singleton or duplicate) were measured for 25(OH)D_3 and 25(OH)D_2 using JCTLM-accepted LC-MS/MS RMPs. Protocols were repeated on 3–4 occasions to generate campaign results. We used separate zeta-score acceptability criteria for daily (|2|) and campaign (|1|) evaluations. Allowable imprecision was determined experimentally.

Results: Across 7 campaigns, unacceptable daily zeta-scores required repeating 2 runs for 25(OH)D_3 and 5 runs for 25(OH)D_2. Hence, the zeta-scores of acceptable reference material results indicated high accuracy. The allowable imprecision for the RMPs was 5% (daily) and 3% (campaign) for 25(OH)D_3 and 7% (daily) and 4% (campaign) for 25(OH)D_2, respectively.

Conclusions: Using zeta-scores and experimentally derived imprecision, we developed a straightforward approach to assess the acceptability of individual 25(OH)D reference measurements, providing also much-needed practical accuracy specifications for 25(OH)D_2.

Keywords

Vitamin D; 25(OH)D_3; 25(OH)D_2; quality performance specifications; serum; reference measurement procedure

1. Introduction

Vitamin D status in humans is currently assessed by measuring the serum concentrations of two liver metabolites, 25-hydroxyvitamin D_3 and 25-hydroxyvitamin D_2 [25(OH)D_3 and

[*]Corresponding Author: Rosemary L Schleicher. Division of Laboratory Sciences, National Center for Environmental Health, Centers for Disease Control and Prevention. 4770 Buford Hwy, NE, Mail Stop F55, Atlanta, GA 30341. zwa5@cdc.gov. Phone: 770-488-4424. Fax: 770-488-4139.

25(OH)D$_2$], usually expressed as the sum of both, so-called total 25-hydroxyvitamin D [25(OH)D]. Accurate measurement of both vitamin D metabolites is essential for interpreting patient status. To improve the accuracy and reliability of 25(OH)D laboratory testing, reference measurement systems consisting of reference measurement procedures (RMPs) [1–3] and higher order standard reference materials (SRMs) [4] have been developed. According to metrological traceability, RMPs are used to assign target values to candidate reference materials (RM), which can be used for trueness assessment of lower-order measurement procedures, and ultimately improve the accuracy of routine measurements. Each assigned target value has measurement uncertainty associated with it. Factors known to have a significant effect on a measurement result are included in the uncertainty calculation, e.g., imprecision, calibrator purity, sample preparation effects of unspecific interferences, weight and density measurements [3].

We developed two liquid chromatography-tandem mass spectrometry (LC-MS/MS) RMPs for quantitation of serum 25(OH)D$_3$ and 25(OH)D$_2$, recognized by the Joint Committee for Traceability in Laboratory Medicine (JCTLM) [5], and currently used to support the CDC Vitamin D Standardization Certification Program [6]. This program was developed to improve the accuracy of 25(OH)D testing and focuses on the enrollment of manufacturers of test kits to provide wide-reaching benefits to kit users.

To provide highly accurate measurements and assure analytical data quality, RMPs must have strict and clearly defined analytical accuracy specifications. Such quality performance goals have been proposed for accuracy evaluation of reference measurements of total 25(OH)D, namely bias 1.7%, and imprecision CV 5% [7, 8]. We performed method validation using these pre-defined quality specifications, because they are considered to be a good balance between a state-of-the-art routine LC-MS/MS method and the required accuracy for a fit-for-purpose RMP [3]. These quality performance goals were developed from different data sources, specifically, from 25(OH)D biological variation and the imprecision of state-of-the art routine methods. It has been suggested that the bias and imprecision goals could be applied to the individual forms, 25(OH)D$_3$ and 25(OH)D$_2$, if they occur in equimolar amounts [7]. However, in human serum the concentration of 25(OH)D$_3$ and 25(OH)D$_2$ are typically very different from one another. Vitamin D$_3$, the precursor of 25(OH)D$_3$ is produced by action of sunlight on skin and obtained from diet and over-the-counter supplements. In the U.S., vitamin D$_2$ is primarily obtained through prescribed supplementation and therefore serum 25(OH)D$_2$ is typically present at very low concentrations or not present at all, compared to the major metabolite 25(OH)D$_3$. Trying to use the suggested quality performance specifications for higher-order measurements for both metabolites poses problems because of the vastly different concentrations in serum.

Because the current literature lacks accuracy rules for individual vitamin D metabolites and does not provide clear guidance on how to assess a measured concentration against an established target concentration when multiple measurements are involved, we developed in-house criteria along with procedures that we applied to single and multiple runs to assess analytical accuracy. We developed these criteria based on the years of experience running the reference method in our laboratory. Our approach features accuracy evaluations of multiple RMs analyzed together with candidate RM (unknowns) over multiple days. This

paper presents our in-house analytical quality specifications for a single measurement series (run) and for multiple independent reference measurements series (campaign, 3 runs).

2. Materials and methods

Reference standard for 25(OH)D$_3$ was purchased from U.S. Pharmacopeia Convention (USP, Rockville, MD). Isotopically labeled 25-hydroxyvitamin D$_2$-[^{2}H$_3$] (d_3-25(OH)D$_2$) and 25-hydroxyvitamin D$_3$-[^{2}H$_6$] (d_6-25(OH)D$_3$) were purchased from Medical Isotopes (Pelham, NH). Reference materials, SRM 2972a and SRM 972a, were purchased from the National Instate of Standards and Technology (NIST, Gaithersburg, MD). ACS grade methanol and *n*-hexane were obtained from Burdick & Jackson (Muskegon, MI). ASC/USP grade ethanol was purchased from Pharmco-AAPER (Brookfield, CT). Purified water (18 MΩ) was obtained from a water purification system (Aqua Solutions, Inc., Falmouth, ME). In-house developed reference materials (Ghent RM) were prepared from pooled human serum, obtained from anonymous blood donors (Tennessee Blood Services, Memphis, TN). The analysis of these RMs by the Centers for Disease Control and Prevention (CDC) laboratory was determined not to constitute engagement in human subject research. For these in-house Ghent RMs the concentration of each 25(OH)D metabolite was assigned by JCTLM-accepted reference methods at Ghent University (Belgium). Captiva 96-well filter plates (0.45 μm pore size) were purchased from Agilent (Santa Clara, CA). The two analytical columns, used in the reported analytical methods [3, 9] were the Ascentis Express F5 (2.1 mm x 150 mm x 2.7 μm), Sigma-Aldrich (St. Louis, MO) and the Acquity HSS PFP (3.0 mm x 150 mm x 1.8 μm), Waters (Milford, MA).

2.1. Calibration

Calibration of our RMPs was based on a previously described procedure [3] with modifications described in this section, namely, choice of calibration matrices (serum or solvent) and a narrower mass ratio range. Individual working internal standard solutions (ISTD) were prepared from concentrated internal standard stock solutions (prepared from solid material) with a targeted mass concentration of 25 ng/mL (32 nmol/L) for d_6-25(OH)D$_3$ and 4 ng/mL (10 nmol/L) for d_3-25(OH)D$_2$; the exact mass concentrations were determined at the time of gravimetric preparation. For solvent-based calibration, 2 independent ethanolic working solutions were gravimetrically prepared from SRM 2972a [10] using calibrated balances under environmentally controlled conditions. The targeted mass concentration was 40 ng/mL (100 nmol/L) for 25(OH)D$_3$ and 8 ng/mL (20 nmol/L) for 25(OH)D$_2$; the exact mass concentrations were determined at the time of gravimetric preparation. Alternatively, for serum-based calibration, we used certified reference materials, e.g., SRM 972a [11] or RM prepared in-house from pooled human serum with concentrations provided by the JCTLM-accepted RMPs at Ghent University [2]. For each analyte, 2 independent calibration curves (3 calibration levels each) were prepared, 1 from each independent ethanolic working solution (solvent-based) or from 2 different RMs (serum-based). To prepare each calibration level, we gravimetrically added individual working ISTD solutions (0.500 mL) to a pre-calculated amount (0.125–0.555 mL) of serum- or solvent-based calibration level to obtain a mass ratios of unlabeled to labeled analyte (analyte/ISTD) of approximately 0.7, 1.0, and, 1.3. Solvent-based calibrators (containing the

standard and ISTD) were evaporated (in a vacuum centrifuge at 45 °C or at room temperature under N_2), reconstituted in 75% methanol/water, and injected in the instrument. For serum-based calibration, 1 mL of deionized water was added to each calibrator (containing the serum RM and ISTD). After 1 hour (room temperature, dark) we adjusted the pH to ~10 with 0.1 g/mL Na_2CO_3 (200 μL) to release the metabolites from vitamin D binding protein. After thorough mixing, the analytes of interest and their ISTDs were extracted twice with hexane, following the previously described procedure for liquid-liquid extraction from serum [3]. We evaporated the combined extracts (vacuum, 45 °C), followed by reconstitution with 75% methanol/water (0.300 mL). The serum extracts were filtered (0.45 μm filter plate) and injected for LC-MS/MS analysis. To assess the trueness of measurement for each calibration matrix, we analyzed 2 Ghent RMs (RM 001 and RM 003) using solvent- and serum-based calibrations in 3 independent campaigns. The assigned reference mass concentrations for RM 001 and RM 003 were 16.3 (40.7) and 11.8 (29.5) ng/mL (nmol/L) for 25(OH)D$_3$ and 1.24 (3.0) and 5.43 (13.2) ng/mL (nmol/L) for 25(OH)D$_2$, respectively.

For the routine LC-MS/MS procedure, we used serum-based calibrators prepared similarly to previously described calibrators in PBS-4% albumin [9]. We prepared 7 calibration levels by mixing ethanolic stock solutions of 25(OH)D$_3$ and 25(OH)D$_2$ with low analyte baseline serum. The concentration range of the calibration levels was from 2.9 to 57.7 ng/mL (7.2 to 144 nmol/L) and from 0.9 ng/mL to 24.7 ng/mL (2.2 to 60 nmol/L) for 25(OH)D$_3$ and 25(OH)D$_2$ respectively. For samples with high concentrations, we used one additional high calibrator per analyte at 117 and 47 ng/mL (292 and 114 nmol/L), respectively. The value assignments for 25(OH)D$_3$ and 25(OH)D$_2$ calibrators were confirmed by the CDC RMPs. ISTD solutions in 67% ethanol/water were used at 30 ng/mL (75 nmol/L) d_6-25(OH)D$_3$ and 8 ng/mL (19 nmol/L) d_3-25(OH)D$_2$.

2.2 Sample preparation

All samples for 25(OH)D used in our RMPs were prepared according to a previously described procedure for liquid-liquid extraction from serum, with the modifications described in this section, namely a mass ratio of approximately 1 (analyte/ISTD) [3]. Each sample was spiked gravimetrically with pre-determined amount of ISTDs to get approximately a 1:1 mass ratio of analyte to ISTD for each analyte; wider mass ratios were used previously. All samples were initially screened for 25(OH)D$_3$ and 25(OH)D$_2$ using our routine method to obtain orientation values for each analyte. We gravimetrically added pre-calculated amounts of isotopically labeled ISTDs to serum (0.200–0.750 g, typically 0.500 g). After 1 h equilibration with the ISTD, the pH was adjusted (0.2 mL of 0.1 g/mL Na_2CO_3), the samples were mixed, and the analytes of interest were extracted with hexanes, using a previously described procedure [3]. Our RMP protocol included in each independent measurement series (run) the analysis of 1 NIST RM in singleton and 2 different Ghent RMs and all CDC candidate RMs (unknowns, typically 10) in duplicate. We carried the analysis of this set of samples typically through 3 runs, which provided the final campaign result, as a mean of all daily measurements for each sample. A diagram of the process is presented in Figure 1.

For routine measurements, we mixed 0.100 mL of serum sample with ISTD solution (0.075 mL) and 72% MeOH/water (0.1 mL) followed by liquid-liquid extraction with hexanes (1.5 mL) according to a previously published procedure [9]. We analyzed all samples used for imprecision studies with the RMP design protocol shown in Fig. 1.

2.3. LC-MS/MS methods

The RMPs featured isocratic (Ascetis Express HPLC column) [3] or alternatively gradient chromatographic separation (Acquity HPLC column) with atmospheric pressure chemical ionization (APCI) in positive ion mode and tandem mass-spectrometric detection. The routine method was isotope dilution LC-MS/MS with isocratic separation and positive APCI ionization detection mode [9].

2.4. Precision

We used data generated with our RMPs for RMs and CDC candidate RMs to develop quality specifications for imprecision. We selected 20 specimens each for $25(OH)D_3$ (no specific criteria regarding concentration) and $25(OH)D_2$ (targeting samples with mass concentrations of $25(OH)D_2$ above the limit of quantitation, LOQ (0.6 ng/mL, 1.5 nmol/L) and below 15 ng/mL (36 nmol/L)). The mass concentrations in the study samples ranged from 9.0 to 54 ng/mL (from 22.5 to 135 nmol/L) for $25(OH)D_3$ and from 0.64 to 13.6 ng/mL (from 1.6 to 33 nmol/L) for $25(OH)D_2$. Mass concentrations (X_1 and X_2) from daily duplicate preparation were averaged to obtain the daily mass concentration (X_D); to assess the daily variability we calculated the relative pair difference as a percent [abs($X1-X2$)/X_D x 100] (Fig. 1). Daily mass concentrations over 3 days were used to calculate the campaign mass concentration (X_C) for each material. We calculated the campaign imprecision CV_C from the daily mass concentrations (X_D). We also selected 20 specimens from the routine LC-MS/MS method, using similar concentration requirements for $25(OH)D_2$ (above the limit of detection, LOD (0.84 ng/mL, 2.0 nmol/L) to assess imprecision compared to the RMPs. The 20 specimens were not the same across methods or analytes. The mass concentrations in the study samples ranged from 6.8 to 41 ng/mL (17 to 102 nmol/L) for $25(OH)D_3$ and from 0.89 to 8.7 ng/mL (2.2 to 21 nmol/L) for $25(OH)D_2$. We calculated the mean and median relative pair differences from all daily relative pair differences (n=60) and campaign CV_C (n=20) for each analyte from reference and routine measurements.

We then applied the quality performance specifications for imprecision to reference data for CDC candidate RMs obtained during 3 recent campaigns, which we will call "training set." The training set included 31 samples for $25(OH)D_3$ (4.8–42 ng/mL, 12–105 nmol/L); only 9 of these samples had reportable concentrations for $25(OH)D_2$ (0.62–14.2 ng/mL, 1.5–34 nmol/L).

2.5. Accuracy

Typically we analyze at least 1 NIST RM material, e.g., SRM 972a (prepared in singleton) [11], and at least 2 Ghent RMs (prepared in duplicate) with each reference run. The detailed data evaluation procedure is outlined in Table 1. Typically if one or multiple CDC candidate RM violate the imprecision specification, those reference measurement results are rejected and the samples are re-analyzed in another measurement series (repeat CDC candidate RM).

If one Ghent RM or the NIST SRM violate the accuracy or imprecision criteria, the reference run is rejected and all samples are re-analyzed in another measurement series (repeat run). We used data from reference measurements of 2 materials, namely, SRM 972a level 2 and RM 003 from 7 recent campaigns, to demonstrate the accuracy of our RMPs. We selected these 2 RMs because both of them were analyzed in these campaigns and because the target concentrations were assigned by different reference methods, NIST and Ghent University.

We opted to use zeta-scores (ζ) for quantitative assessment of accuracy. This standardized parameter shows how well a measurement matches the assigned target concentration relative to the combined uncertainty of the measurement and the target concentration [12, 13]. A zeta-score of 0 indicates a perfect match. In proficiency testing programs, a zeta-score of ± 2 indicates acceptable participant result. The zeta-score is defined as follows:

$$\zeta = \left(x - y \right) / \sqrt{u(x)^2 + u(y)^2} \qquad \text{(equation 1),}$$

where x is the CDC RMP result and y is the NIST or Ghent RMP result; $u(x)$ is the standard uncertainty of the CDC RMP [3] and $u(y)$ is the reported standard uncertainty for the Ghent and NIST RMs. For SRM 972a level 2, we used the target concentrations (y) listed in the Certificate of Analysis (COA) and the reported standard uncertainties $u(y)$ of 1.1% and 3.8% for 25(OH)D$_3$ and 25(OH)D$_2$, respectively [11]. The standard uncertainties were derived by dividing the reported expanded uncertainty [$U=2\,u(x)$] from the certificate by the coverage factor k=2. For RM 003, we used the target concentrations (y) of 11.8 ng/mL (29.5 nmol/L) for 25(OH)D$_3$ and 5.43 ng/mL (13.2 nmol/L) for 25(OH)D$_2$ and the reported standard uncertainties $u(y)$ of 1.5% and 1.8% for 25(OH)D$_3$ and 25(OH)D$_2$, respectively, as provided in the report of analysis. The estimated standard uncertainties $u(x)$ for CDC RMPs were 1.6% and 1.8% for 25(OH)D$_3$ and 25(OH)D$_2$, respectively [3].

To evaluate the daily accuracy, we calculated the zeta-score$_D$ as described above, using the certified result and uncertainties from the NIST COA for the SRM material [11], and the Ghent assigned target value and reported uncertainties for the Gent RM material, listed above. In both cases, we used our daily reference measurement (x: X$_D$) and the uncertainty $u(x)$ for our RMP. To evaluate the campaign accuracy, we calculated the zeta-score$_C$ from the campaign mass concentration (x: X$_C$). Similarly, we determined the daily and campaign systematic deviation from the target concentration (bias, %) for each material. We developed a macro in SAS version 9.3 to evaluate daily and campaign reference measurements and to automatically apply the numerical quality performance specifications for accuracy.

Results and discussion

3.1. Calibration

In our laboratory, we have well-established methods for analysis of serum 25(OH)D metabolites, namely, LC-MS/MS routine and reference measurement procedures [9, 3]. For the RMPs, we studied 2 types of calibrations, solvent- and serum-based. For solvent-based calibration, ethanolic working solutions were prepared by gravimetric dilution of SRM

2972a with absolute ethanol. This is the most direct way to establish traceability to the highest order of commercially available accuracy-based NIST RM and SI units. All working solutions were stable when stored in tightly closed glass containers at −20 °C ± 2 and used within 1 month of preparation. To prepare serum-based calibrators, we used SRM 972a level 1 or level 4 for 25(OH)D$_3$ and SRM 972a level 3 for 25(OH)D$_2$. These 3 levels were chosen because they had the highest certified concentrations for each respective analyte, and therefore 1 vial each was sufficient to prepare 1 independent calibration curve. The second calibration curve for each analyte was prepared from Ghent RMs. We used both curves for quantitation. To compare the calibration matrices (solvent vs. serum) with regards to method trueness, 2 RMs (RM 001 and RM 003) were processed independently with each calibration matrix in 3 independent campaigns. We evaluated the calculated mass concentrations from each curve against the target concentrations for accuracy, expressed quantitatively with the zeta-score. Using 2 RMs over 3 campaigns, the mean daily zeta-score$_D$ (SD) from serum- vs. solvent-based calibrations was −0.93 (0.66) vs. 0.01 (0.94) for 25(OH)D$_3$ and 0.43 (1.15) vs. 0.11 (1.25) for 25(OH)D$_2$, respectively. The overall daily bias from the serum-based curve was −2% and 1% for 25(OH)D$_3$ and 25(OH)D$_2$, respectively. The overall daily bias from the solvent-based curve was −0.1% and 0.3% for 25(OH)D$_3$ and 25(OH)D$_2$, respectively. Accuracy from solvent-based calibration showed nearly perfect agreement, as expressed by a mean zeta-score of approximately 0 and a mean bias 0.3%. Our preferred RMP calibration approach is therefore solvent-based.

3.2. Sample preparation

For reference measurements, all samples were prepared by accurate weighing of serum and pre-calculated amounts of isotopically labeled ISTD solutions to obtain approximately 1:1 mass ratios of analyte to ISTD. In our earlier studies, the amount of working ISTD was not closely matched to the amount of measurand, resulting in wide mass ratios, 0.25–2.50. The impact of the use of the different mass ratios (wide vs narrow) on overall accuracy of all valid results has not been shown to be significantly different in our laboratory (data not shown). We noted improvement with imprecision violations using the narrow mass ratio from 4% and 17% for 25(OH)D$_3$ and 25(OH)D$_2$ down to 0% and 10%, respectively. Daily zeta-score violations were marginally improved for 25(OH)D$_3$ (3% to 2%) and greatly improved for 25(OH)D$_2$ (18% to 7%). Considering the low throughput (10 samples/run) and laborious sample preparation for RMPs, switching to a mass ratio of analyte to ISTD of approximately 1:1 was beneficial.

3.3. Analytical quality performance specifications

During the development and validation of the RMPs [3] we used the pre-defined proposed quality specifications for accuracy, i.e., CV<5% and bias <1.7% [7]. Our RMPs showed similar performance characteristics to those of other established RMPs [1, 2]. We used our RMPs to analyze the same sample over multiple days (campaign), which involves complex data review with independent daily and campaign data assessment from multiple RM. Because the accuracy specifications for total 25(OH)D are not applicable to the individual metabolites occurring at very different concentrations, we developed our in-house numerical quality performance imprecision and accuracy goals, based on years of experience with the method. We had to consider the complex multiday data analysis protocol of our RMPs (Fig.

1). We evaluated all daily data for imprecision and accuracy. If the run passed our criteria, the exact same analyses were repeated on 2 more days and the same daily data review was conducted. Data from all 3 valid runs comprise the campaign data; the combined data were evaluated for overall campaign imprecision and accuracy. For a summary of our quality performance specifications procedure, acceptability criteria, and actions taken upon data evaluation, see Table 1.

3.3.1.   Daily precision—In developing numerical quality performance goals for daily imprecision evaluation, we used data from the CDC RMPs for 20 specimens analyzed in duplicate over 3 days. Even though the $25(OH)D_2$ mass concentration for the study samples covered a wide range (0.64–13.6 ng/mL, (1.6–33 nmol/L), the majority of the samples had a mass concentration <3 ng/mL (7 nmol/L). We assessed the relative pair difference as a measure of daily variance. Plots of the daily relative pair differences as a function of the daily mass concentration (X_D, ng/mL) from all daily reference measurements (3 days, duplicate preparations, n=60) are presented in Figure 2 for each metabolite (panels A and B), where the median relative pair difference and the 95th percentile are depicted by solid and dotted lines, respectively. The plots for both analytes showed equally distributed variability independent of the concentration. The distribution of the relative pair difference for the major metabolite was slightly right skewed (not shown) and therefore we calculated the median in addition to the mean and we chose to use the 95th percentile of the relative pair differences to set the daily imprecision limit for each metabolite (Table 2). The median relative pair difference (n=60) was 2.1% for $25(OH)D_3$ and 3.0% for $25(OH)D_2$ and our daily imprecision limit was   5% (rounded from 5.2%) for $25(OH)D_3$ and   7% for $25(OH)D_2$.

To confirm the daily imprecision goals, we used a common approach that compared the performance of the RMPs to that of a hierarchically lower measurement procedure, namely, a routine method. The expectation is that the overall measured imprecision of the RMPs should be half of that of a routine method LC-MS/MS [7, 8]. We calculated the relative pair differences for each analyte for the selected study samples from routine measurements and plotted it as a function of the daily mass concentration, X_D (Fig. 2 panels C and D). Plot D in Figure 2 shows increased variability at lower mass concentrations for $25(OH)D_2$, however, the majority of measurements with higher variability were at concentrations lower than the LOQ of the assay 2.5 ng/mL (6.1 nmol/L) with the maximum relative pair difference of 24% recorded at 0.96 ng/mL (2.3 nmol/L). The median relative pair difference of the routine method (n=60) depicted with a solid line, was 4% for $25(OH)D_3$ and 5% for $25(OH)D_2$. The 95th percentile, depicted with a dotted line, was 11% for $25(OH)D_3$ and 21% for $25(OH)D_2$. Thus, the median relative pair difference for the RMPs were close to half of that obtained for the routine measurements for both analytes (Table 2). Our lower order LC-MS/MS method for serum 25(OH)D measurements [9] historically featured excellent long-term imprecision (CV of 3% for $25(OH)D_3$ and 5% for $25(OH)D_2$) [14]. In a recent review of 24 LC-MS/MS procedures the inter-assay imprecision was listed for 13 routine LC-MS/MS methods, i.e., mean CV of 7.1% for $25(OH)D_3$ and 7.8% for $25(OH)D_2$ [15]. The reported imprecision of our routine method was in line with and even superior to

the imprecision performance reported by other routine LS-MS/MS 25(OH)D procedures and thus can be appropriately used to confirm the daily imprecision goals.

Had we derived cutoffs for reference imprecision from our routine measurements, the limits would have been 5.5% for 25(OH)D_3 and 10.5% for 25(OH)D_2 (half of the calculated 95th percentile). Our RMP imprecision limits for 25(OH)D_3 and 25(OH)D_2 (5% and 7%, respectively) derived from reference data more than met these theoretical goals from routine data.

We applied our daily quality performance imprecision limits to all CDC candidate reference materials from the "training set" analyzed with CDC RMPs. For all RMs and CDC candidate RMs the daily imprecision goals were met 100% for 25(OH)D_3 and 25(OH)D_2. The mean relative pair difference was 1.5% for 25(OH)D_3 and 3.1% for 25(OH)D_2. The maximum relative pair difference was 4.4% and 6.1% for 25(OH)D_3 and 25(OH)D_2, respectively. The 95th percentile for this recent set of data (n=31) was 4.0% for 25(OH)D_3 and 6.0% for 25(OH)D_2, which suggests that our daily difference limits may be reduced to 4% and 6%, respectively. The daily mean relative pair differences were less than half of the limits, nearly ensuring that the maximum goals were not exceeded. This indicates that our RMPs consistently met the daily imprecision performance goals, despite the fact that most of the 25(OH)D_2-containing samples were below 3 ng/mL (7.3 nmol/L).

3.3.2. Campaign precision—We developed numerical quality performance specifications for campaign imprecision reference measurements using the same study set of 20 specimens, analyzed over 3 days with our RMPs. For each specimen, the mass concentration from independent daily measurement (X_D) was used to calculate the campaign imprecision, CV_C as depicted in Figure 1. The distribution of the CV_C as a function of the mass concentration (X_C) for each of the 20 serum specimens is illustrated in Figure 3 (**panels A and B** for reference measurements and **panels C and D** for routine measurements). Similarly to the daily trends, the campaign plot for 25(OH)D_2 (Figure 3D) showed a marked increase in imprecision at low mass concentrations (<3 ng/mL (7.3 nmol/L) for routine measurements, while the RMP data showed a slight trend towards higher imprecision with lower concentrations (Figure 3B). We used the same concept to establish the campaign imprecision limits based on the calculated 95th percentile of the CV_C distribution from RMPs: 3% for 25(OH)D_3 (rounded from 3.1%) and 4% for 25(OH)D_2 (rounded from 4.2%) (Table 2). The calculated 95th percentile CV_C from routine measurements was at least twice that of reference measurements for each analyte, which demonstrated the validity of the selected cut-offs, confirmed by running samples with the routine method. For all RMs and CDC candidate RM from the "training set" the campaign imprecision goals were met 100% for both analytes. The mean CV_C was 0.9% for 25(OH)D_3 and 2.0% for 25(OH)D_2, which were at least half of the campaign cut-offs. Thus, the RMPs consistently met the campaign imprecision quality performance goals, which were stricter than the proposed specifications of 5%. Based on all of the above, we can claim with confidence that our numerical imprecision goals for campaign reference measurements fit the purpose of intended use.

3.3.3.   Daily accuracy—The result from multiple serum reference materials, a single preparation of NIST RM (X_D, ng/g) and the mean result from duplicate preparations of each Ghent RM (X_D, ng/g), were evaluated against the certified mass fraction concentration of each analyte and RM for equivalence. We looked into approaches for quantitative assessment of the equivalence between 2 measurement results of the same RM, obtained by different laboratories.

A panel of experts recommended using the bias approach for total 25(OH)D, which we followed during the method validation process [7, 8]. However, the minor metabolite, 25(OH)D$_2$ is generally undetectable or just above the limit of detection in the US population [14]. Random and systematic effects can easily affect measurements at low concentrations. The factors that are likely to have a significant effect on a measurement result, based on our long experience performing vitamin D measurements, are included in the uncertainty, e.g., type A errors such as imprecision and type B errors such as calibrator purity, sample preparation effect of unspecific interferences, or weight and density measurements. The type A variance is calculated from the variability of repeated measurements, which is more notable at lower concentrations, resulting in higher uncertainties [3]. In the COA for SRM 972a, NIST lists the certified value (Y) for each analyte in the certified matrix as $Y_{NIST} \pm U_{95}(Y_{NIST})$; this means that the interval $Y_{NIST} - U_{95}(Y_{NIST})$ to $Y_{NIST} + U_{95}(Y_{NIST})$ is expected to contain the true value of Y with a 95% level of confidence. For example, NIST offers 2 materials with certified 25(OH)D$_2$ mass concentrations for SRM 972a, at 0.81 (2.0) and at 13.3 (32.3) ng/mL (nmol/L), where the expanded uncertainty associated with each target is 7.4% and 2.3%, respectively [11]. Thus, at the lower mass concentration (0.81 ng/mL, 2.0 nmol/L), the uncertainty is significantly higher. When using the so called "acceptance interval" (target value $\pm U_{95}$) approach with SRM 972a [11], we can assess how well our "reference measurement $\pm U_{95}$" (calculated range) complies with the certified target interval.

Next, we looked for guidance from procedures used by proficiency testing (PT) programs to assess the difference between the participant's result and the assigned value. In a review article for scoring results in PT programs, Miller *et al.* pointed out that when the acceptance interval is expressed as a percent (e.g., ±15%), the concept may be not be reasonable below a certain concentration, because the SD of a measurement procedure becomes a larger fraction of the acceptance interval. To overcome this problem, the authors proposed to use a fixed unit interval concept (e.g., $\pm$ XX ng/mL (nmol/L)) instead of a percentage [16]. Both of these assessment approaches, involving "acceptance intervals" could be easily adopted, however, we wanted to find a single measure to assess how well CDC's reference measurement matched a target result.

International guidelines and a panel of experts from the Royal Society of Chemistry's Analytical Methods Committee explained in detail how a zeta-score could be used in PT programs as a standardized quantitative indicator to assess accuracy of the participant's result [12, 13]. The zeta-score calculation incorporates the uncertainties associated with each measurement procedure [participant's method (x) and target assignment method (y)]. For meaningful scoring, the 2 measurements (x and y) should have similar and relatively low uncertainties. The standard uncertainties of the CDC RMPs are comparable with the

uncertainties reported by Ghent and NIST. In our hands, the zeta-score is a good accuracy indicator at all concentrations. For daily reference measurements, we set a zeta-score$_D$ of ±2 as acceptable (corresponding to a 95% confidence level for 2-sided confidence intervals assuming a normal distribution, or 2SD). We believe that this cut-off point fits the purpose of use and will allow us to make a technically correct decision about the validity of the run.

In this report, we demonstrate the daily accuracy quality performance of the CDC RMPs using 2 serum reference materials, namely, SRM 972a level 2 and RM 003 analyzed in 7 campaigns. We selected these 2 RMs because they had notably different target concentration for 25(OH)D$_2$, 0.81 and 5.43 ng/mL (1.96 and 13.2 nmol/L), respectively. For the major metabolite, the average zeta-score$_D$ and bias (%) from all daily measurements were 0.1 and 0.1% in SRM 972a-2 and 0.2 and 0.4% in RM 003, respectively (Figure 4, **panels A and B**). For the minor metabolite, the average daily zeta-score$_D$ and bias (%) were −0.2 and −0.9% for SRM 972a and 0.4 and 1.0% for RM 003, respectively (Figure 4, **panels C and D**). In these 7 campaigns, daily negative and positive zeta-scores$_D$ were in near perfect agreement with the daily bias for both metabolites (r^2>0.999, data not shown). A zeta-score of 2 corresponds to a bias of 4% for 25(OH)D$_3$ and between 5 and 8% for 25(OH)D$_2$, depending on the analyte concentration. The low certified reference concentration for 25(OH)D$_2$ and the larger expanded uncertainty respectively in SRM 972a-2 contributed to the overall higher bias of 8% corresponding to a zeta-score of 2. It should be pointed out that to assess measurement accuracy for SRM materials, the zeta-score$_D$ is calculated from a single preparation; nonetheless, the mean bias from all daily reference measurements in the 7 campaigns was 0.1% for 25(OH)D$_3$ and −0.9% for 25(OH)D$_2$ in SRM 972a-2. The majority of zeta-score$_D$ from our RMP measurements were less than half of the upper cut-off point of 2 for both analytes (Figure 4A–D), which indicates excellent accuracy quality performance over a period of 1 year. Accuracy violations (unacceptable daily zeta-scores of 2) across 7 campaigns required repeating 2 runs for 25(OH)D$_3$ and 5 runs for 25(OH)D$_2$. Overall, the accuracy quality performance goal was met in all valid daily reference measurements.

3.3.4.   Campaign accuracy—The campaign zeta-score$_C$ for each metabolite in each RM was calculated from the campaign mass concentration (Xc), according the analysis scheme presented in Figure 1, and used as an indicator for accuracy. Data from 7 campaigns for each metabolite illustrated good correspondence between the calculated campaign zeta-score$_C$ and corresponding bias in each RM (Figure 5). Our cut-off for campaign accuracy was a zeta-score$_C$ of ±1 (corresponding to a 67% confidence level for a 2-sided confidence intervals, or 1SD). The accuracy goal was met for each analyte; the mean zeta-score$_C$ and bias (%) for 25(OH)D$_3$ was 0.1 and 0.1% in SRM 972a-2 and 0.2 and 0.4% in RM 003; and for 25(OH)D$_2$, it was −0.2 and −0.9% in SRM 972a-2 and 0.4 and 1.0% in RM 003 (Figure 5A–5D). Our cut-point for the campaign zeta-score$_C$ fits the purpose of intended use based on the current state-of-the-art, assures the quality of reference measurements, and at the same time is consistent with the proposed specification of maximum systematic deviation for total 25(OH)D of 1.7% [7, 8]. The CDC RMPs consistently achieved the established accuracy goals. The mean of all zeta-score$_C$ and bias (%) across the 7 campaigns and 2 materials (SRM 972a-2 and RM 003) was 0.1 and 0.3% for 25(OH)D$_3$ 0.1 and 0.04% for

$25(OH)D_2$, indicating no bias and confirming the long-term metrological comparability of our RMPs to other RMPs.

4. Conclusions

Our in-house analytical accuracy specifications were tailored to each individual metabolite, $25(OH)D_3$ and $25(OH)D_2$. We used zeta-scores to evaluate accuracy. A maximum zeta-score$_C$ of ± 1, our cut-point for campaign accuracy, was consistently met in 7 campaigns, where the mean zeta-score$_c$ from 2 RMs was nearly 0 for both metabolites, indicating no bias.

For assessment of imprecision of campaign measurements, we developed refined numerical goals from RMP data, namely a maximum campaign imprecision of 3% for $25(OH)D_3$ and 4% for $25(OH)D_2$, which is stricter than the literature goal of 5% CV for $25(OH)D$. We confirmed the selected cutoffs by analyzing samples with our routine LC-MS/MS method. The overall imprecision of the reference method was half of that determined by our state-of-the-art routine measurements, confirming the validity of our numerical quality performance goals. We demonstrated excellent reference method imprecision, where the goals were consistently met in all CDC candidate RM from 3 campaigns, and the mean CV_C was 0.9% for $25(OH)D_3$ and 2.0% for $25(OH)D_2$. The reported daily quality specifications are our in-house limits, developed based on the state-of-the art performance of our RMPs that allow us to systematically check our daily performance and assure that our reference methods essentially achieve the pre-defined quality specifications for total $25(OH)D$. Recent data suggests that the daily imprecision limits may be reduced by 1%. The reported concept may be beneficial to others who are developing reference methods and can be adapted or modified depending on the design and the state-of-the-art technology.

To our knowledge, we are reporting for the first time accuracy specifications for daily and campaign reference measurements tailored for each vitamin D metabolite. Our approach is unique because it only uses data from the RMs analyzed in the same runs as the candidate RMs to independently determine the accuracy during daily runs and campaign assessments. In combination, the reported analytical quality performance goals and detailed application approach provide a unique internal quality control system that is suitable for the intended use and consistently assures high quality reference measurement results.

Acknowledgements

Abbreviations:

25(OH)D	25-hydroxyvitamin D
RMP	reference method procedure
RM	reference material
SRM	standard reference material
JCTLM	Joint Committee for Traceability in Laboratory Medicine
CDC	Centers for Disease Control and Prevention
VDSCP	Vitamin D Standardization Certification Program
NIST	National Institute of Standards and Technology
ISTD	internal standard
LC-MS/MS	liquid chromatography tandem mass spectrometry
ID	isotope dilution
u	standard uncertainty
U	expanded uncertainty
NHANES	National Health and Nutrition Examination Survey

References

[1]. Tai S-C, Bedner M, Phinney KW, Development of a candidate reference measurement procedure for the determination of 25-hydroxyvitamin D_3 and 25-hydroxyvitamin D_2 in human serum using isotope-dilution liquid chromatography-tandem mass spectrometry, Anal. Chem. 82 (2010) 1942–1948. [PubMed: 20136128]

[2]. M Stepman HC, Vanderroost A, van Uytfanghe K, Thienpont LM, Candidate reference measurement procedure for serum 25-hydroxyvitamin D_3 and 25-hydroxyvitamin D_2 by using isotope-dilution liquid chromatography-tandem mass spectrometry, Clin. Chem. 53(3) (2011) 441–448.

[3]. Mineva EM, Schleicher RL, Chaudhary-Webb M, Maw KL, Botelho JC JC, Vesper HW CPfeiffer M, A candidate reference measurement procedure for quantifying serum concentrations of 25-hydroxyvitamin D_3 and 25-hydroxyvitamin D_2 using isotope-dilution liquid chromatography-tandem mass spectrometry, Anal. Bioanl. Chem. 407 (2015) 5615–5624.

[4]. Phinney KW, Bedner M, Tai S-C, Vamathevan VV, Sander LC, Sharpless KE, Wise SA, Yen JH, Schleicher RL, Chaudhary-Webb M, Pfeiffer CM, Betz JM, Coates PM, Picciano MF, Development and certification of a standard reference material for vitamin D metabolites in human serum, Anal. Chem. 84(2) (2012) 956–962. [PubMed: 22141317]

[5]. Joint Committee for Traceability in Laboratory Medicine (JCTLM). http://www.bipm.org/en/committees/jc/jctlm/;JCTLMmethodidentifier:C12RMP2andC12RMP3 (accessed 04 June 2018).

[6]. Centers for Disease Control and Prevention (CDC). Laboratory Quality Assurance and Standardization Programs: Hormone and Vitamin D Standardization Program. Atlanta (GA), CDC, http://www.cdc.gov/labstandards/hs.html (accessed 05 June 2018).

[7]. Stöckl D, Sluss PM, Thienpont LM, Specifications for trueness and precision of a reference measurement system for serum/plasma 25-hydroxyvitamin D analysis, Clin. Chim. Acta 408 (2009) 8–13. [PubMed: 19563791]

[8]. Thienpont LM, Stepman HC, Vesper HW, Standardization of measurements of 25 hydroxyvitamin D_3 and D_2, Scand. J. Clin. Lab. Invest. 72 (Suppl 243) (2012) 41–49.

[9]. Schleicher RL, Encisco SE, Chaudhary-Webb M, Paliakov EM, McCoy LF, Pfeiffer CM, Isotope dilution ultra performance liquid chromatography-tandem mass spectrometry method for simultaneous measurement of 25-hydroxyvitamin D_2, 25-hydroxyvitamin D_3 and 3-epi-25-hydroxyvitamin D_3 in human serum, Clin. Chim. Acta 412 (2011) 1594–1599. [PubMed: 21601563]

[10]. National Institute of Standards and Technology, Certificate of analysis, standard reference material 2972a: 25-Hydroxyvitamin D calibration solutions, NIST, Gaithersburg, 2014.

[11]. National Institute of Standards and Technology, Certificate of analysis, standard reference material 972a: vitamin D metabolites in frozen human serum, NIST, Gaithersburg, 2017.

[12]. Thompson M, Ellison SLR, Wood R, International harmonized protocol for proficiency testing of analytical chemistry laboratories (IUPAC technical report), Pure Appl. Chem. 78(1) (2006) 145–196.

[13]. Royal Society of Chemistry, Analytical Methods Committee, Understanding and acting on scores obtained in proficiency testing schemes, AMC Technical Brief No 11, Dec 2002.

[14]. Schleicher RL, Sternberg MR, Looker AC, Yetley EA, Lacher DA, Sempos CT, Taylor CL, Durazo-Arvizu RA, Maw KL, Chaudhary-Webb M, Johnson CL, Pfeiffer CM, National estimates of serum total 25-hydroxyvitamin D and metabolite concentrations measured by Liquid-chromatography-tandem mass spectrometry in the US population during 2007–2010, J. Nutr. 146 (5) (2016) 1051–1061. [PubMed: 27052537]

[15]. Le Goff C, Cavalier E, Souberbielle J-C, Gonzáles-Antuna A, Delvin E, Measurement of circulating 25-hydroxyvitamin D: A historical review, Practical Lab. Med. 2 (2015) 1–14.

[16]. Miller WG, Jones GRD, Horowitz GL, Weykamp C, Proficiency testing/external quality assessment: current challenges and future directions, Clin. Chem. 57 (12) (2011) 1670–1680. [PubMed: 21965556]

Daily

Campaign
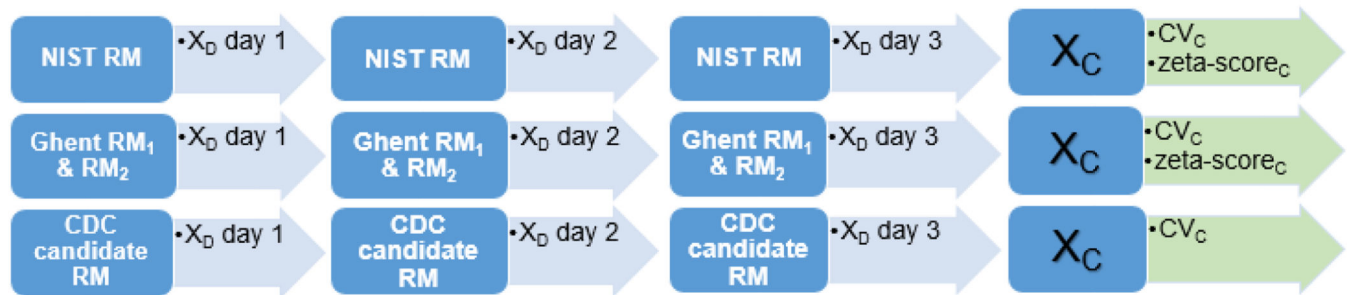
Fig. 1.
Study design of daily and campaign reference measurements. Footnote:*Daily:*RM, reference material prepared in-house and value-assigned at Ghent University's reference laboratoryNIST RM, commercial standard reference materialCDC candidate RM, unknown to be value-assigned using CDC RMPsX_m, mean measurement (ng/mL) from two independent reference measurements, X1 and X2, on the same dayX, single daily reference measurementRelative pair difference, absolute percent difference between X_1 and X_2, divided by the X_mZeta-score$_D$, calculated daily for RM with target values*Campaign:*X_m, mean measurement (ng/mL) from two independent reference measurements, X1 and X2 per dayX_C, mean measurement (ng/mL) from 3 independent measurement seriesX, single daily reference measurementCV_C, mean coefficients of variation for campaign (average from 3 runs)Zeta-score$_C$, calculated for campaign (average from 3 runs) for RM with target values
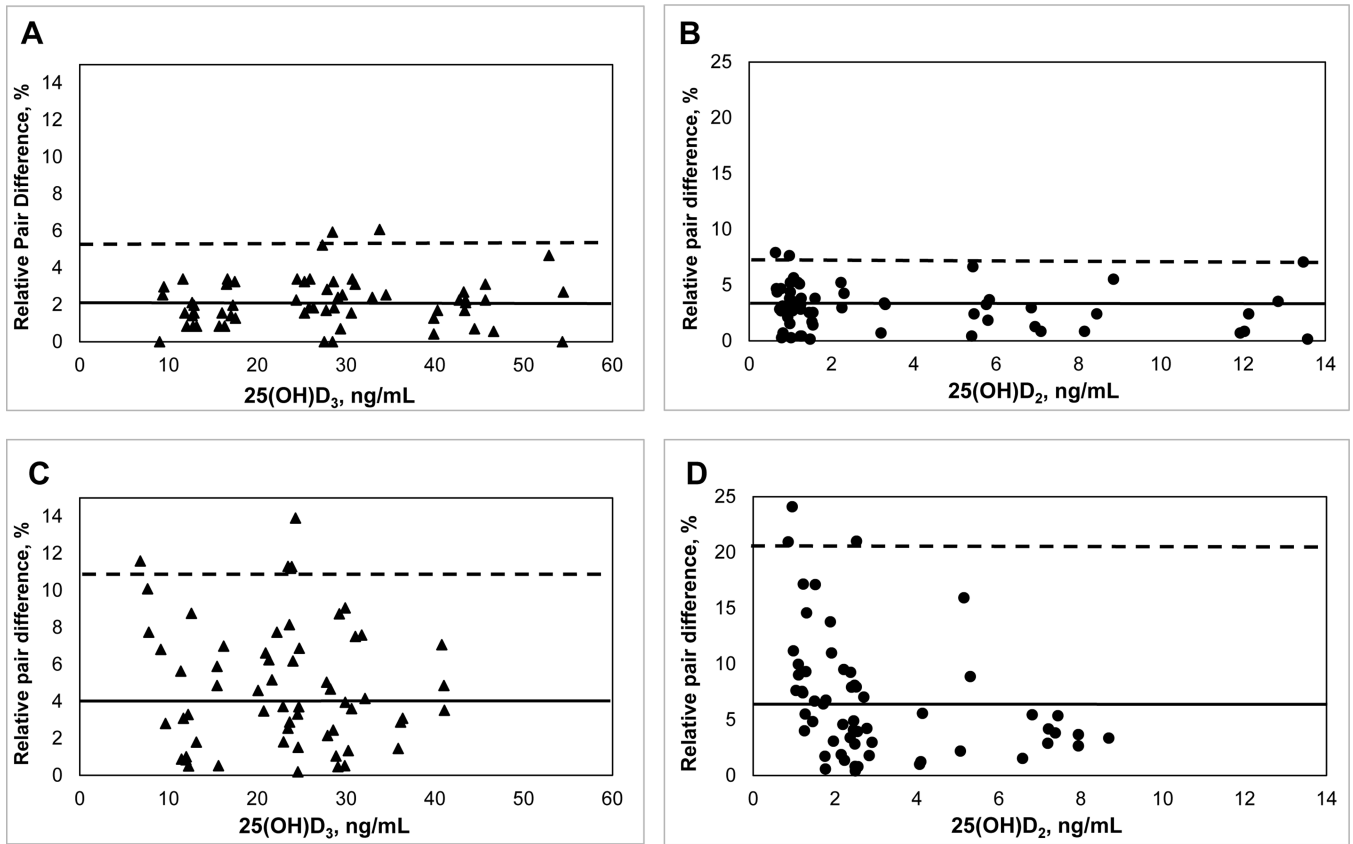
Fig. 2.
Daily 25(OH)D$_3$ and 25(OH)D$_2$ reference (panel A and B) and routine (panel C and D) imprecision (relative pair difference, %) in 20 serum materials (3 days, n=60). Solid line represents the median relative pair difference (%) and dotted line represents the 95th percentile.
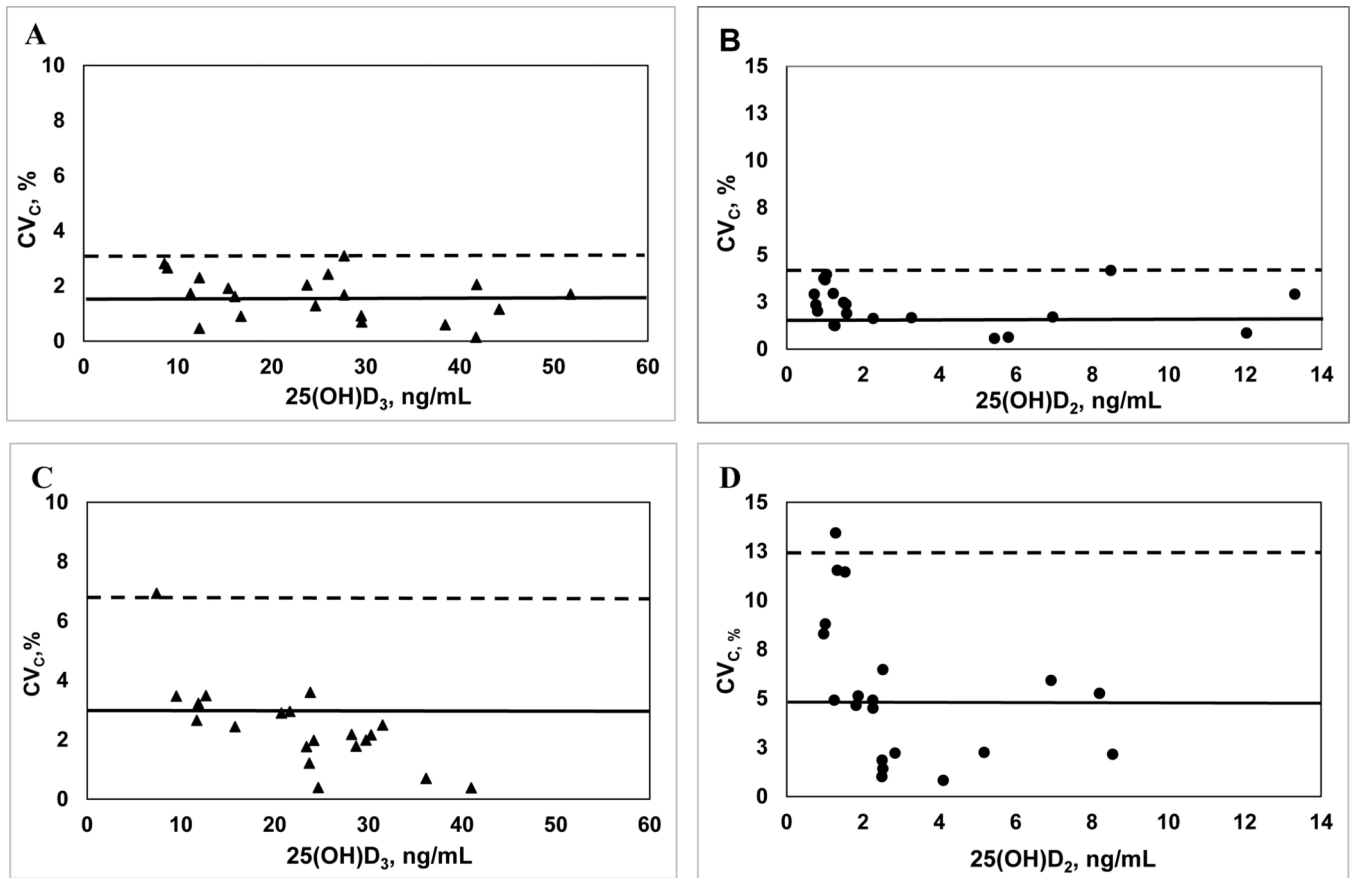
Fig. 3.
Campaign $25(OH)D_3$ and $25(OH)D_2$ imprecision of reference (panel A and B) and routine (panel C and D) in 20 serum materials (n=20). Solid line represents the median CV (%) and dotted line represents the 95th percentile.
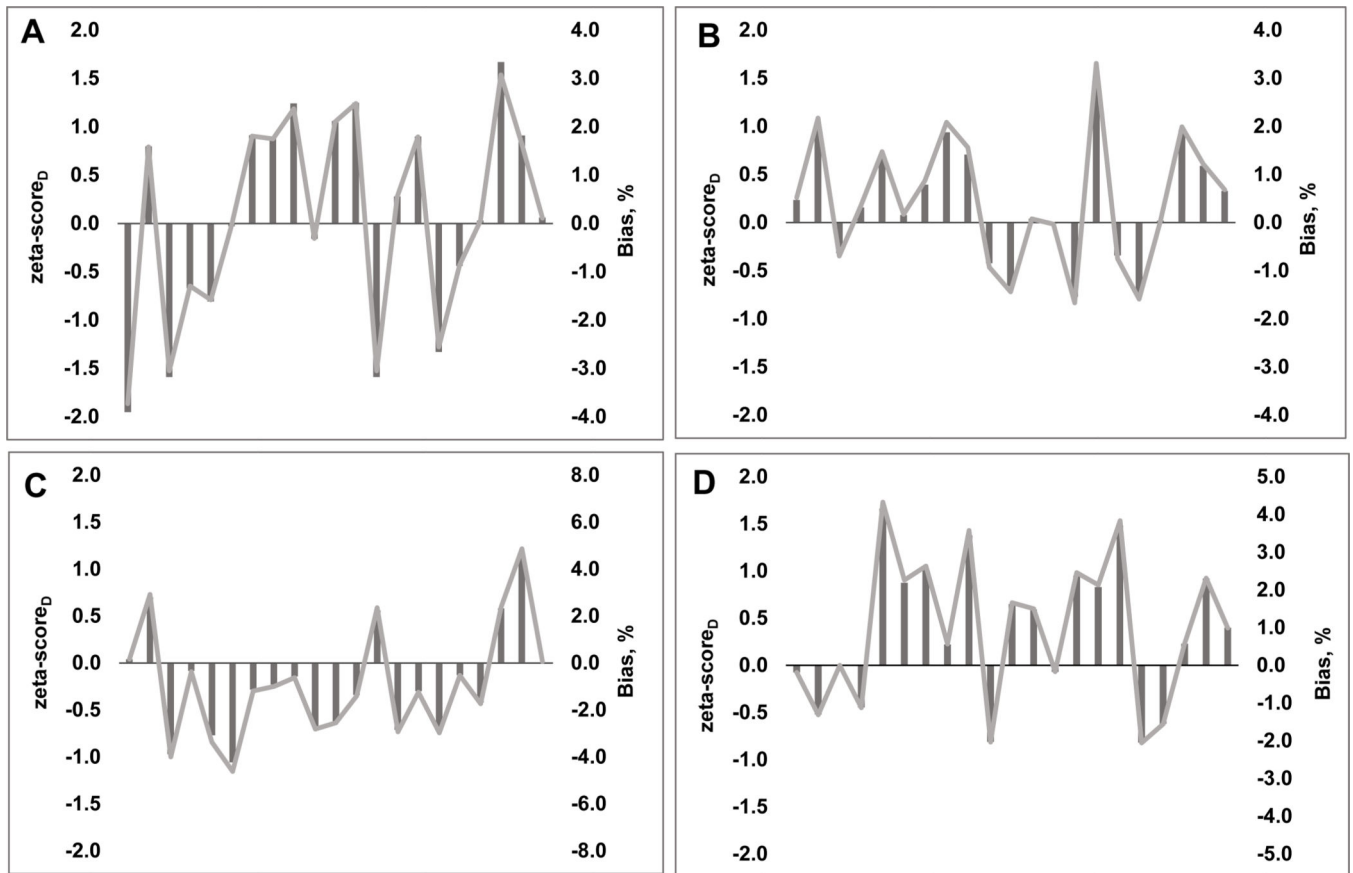
Fig. 4.
Trueness performance (7 campaigns) of CDC's daily reference measurements for 25(OH)D$_3$ (panels A and B) and 25(OH)D$_2$ (panels C and D) using 2 reference materials, SRM 972a-2 (panels A and C) and RM 003 (panels B and D). Footnote: zeta-score: dark gray vertical lines; bias: light gray line.
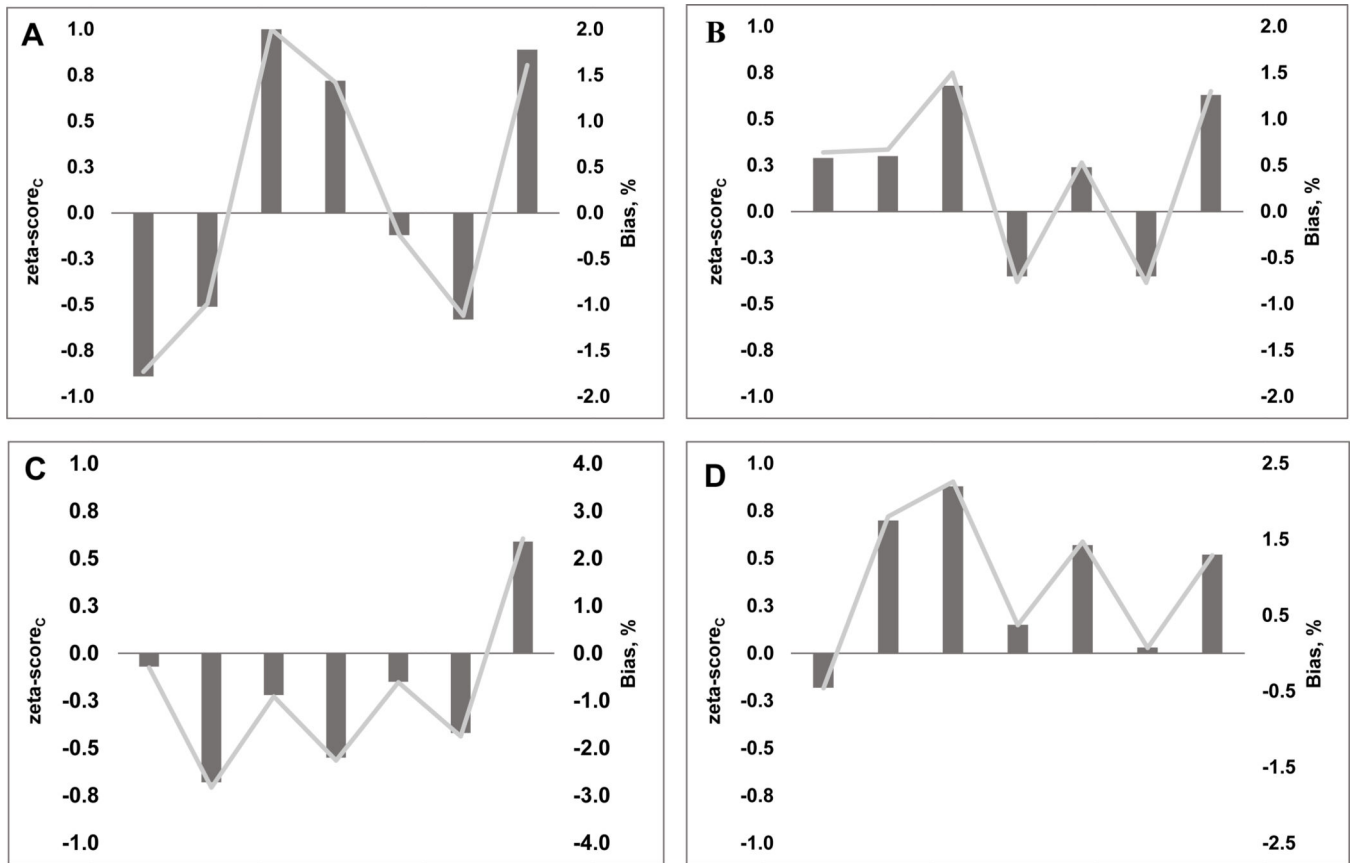
Fig. 5.
Trueness performance (7 campaigns) of CDC's campaign reference measurements for $25(OH)D_3$ (panels A and B) and $25(OH)D_2$ (panels C and D) using 2 RMs, SRM 972a-2 (panels A and C) and RM 003 (panels B and D). Footnote: zeta-score: dark gray vertical lines, bias: light gray line.

Table 1.

Quality performance specifications

Quality parameter	Procedure	Acceptability criteria
Daily imprecision	• Calculate daily relative pair difference from duplicate mass concentration results (X_1 and X_2) for each of 2 RMs for each analyte • Relative pair difference for each RM needs to be within acceptance limit • Repeat run if relative pair difference for ≥1 RM is outside acceptance limit • Repeat CDC candidate RM if relative pair difference is outside acceptance limit	• Experimentally derived as 95th percentile of relative pair difference for 20 specimens analyzed in duplicate over 3 days (n=60) • Relative pair difference ≤5% for 25(OH)D$_3$ and ≤7% for 25(OH)D$_2$ • Criteria applied to concentrations above the LOQ
Campaign imprecision	• Calculate campaign CV_C from 3 daily mass concentration results (X_D) for each of 3 RMs for each analyte • CV_C for each RM needs to be within acceptance limit • Repeat run if CV_C for ≥1 RMs is outside acceptance limit • Repeat CDC candidate RM if CV_C is outside acceptance limit	• Experimentally derived as 95th percentile of CV_C for 20 specimens analyzed in duplicate over 3 days (n=20) • CV_C ≤3% for 25(OH)D$_3$ and ≤4% for 25(OH)D$_2$ • Criteria applied to concentrations above the LOQ
Daily accuracy	• Calculate daily zeta-score$_D$ from daily mass concentration result (X_D) for each of 3 RMs for each analyte • Zeta-score$_D$ for each RM needs to be within acceptance limit • Repeat run if zeta-score$_D$ for ≥1 RMs is outside acceptance limit	• Zeta-score$_D$ ≤|2| for 25(OH)D$_3$ and 25(OH)D$_2$ • Criteria applied to concentrations above the LOQ
Campaign accuracy	• Calculate campaign zeta-score$_C$ from campaign mass concentration result (X_C) for each of 3 RMs for each analyte • Zeta-score$_C$ for each RM needs to be within acceptance limit • Repeat run if zeta-score$_C$ for ≥1 RMs is outside acceptance limit	• Zeta-score$_C$ ≤|1| for 25(OH)D$_3$ and 25(OH)D$_2$ • Criteria applied to concentrations above the LOQ

Table 2.

Daily and campaign imprecision for reference and routine measurements. Each daily estimate consisted of 20 specimens tested in 3 runs.

		Imprecision (%)			
		Reference Method		Routine Method	
		25(OH)D$_3$	25(OH)D$_2$	25(OH)D$_3$	25(OH)D$_2$
Daily	Mean relative pair difference	2.2	3.0	4.8	6.7
	Median relative pair difference	2.1	3.0	4.0	5.3
	95th %ile relative pair difference	**5.2**	**7.0**	11	21
Campaign	Mean CV	1.6	2.2	2.4	5.4
	Median CV	1.7	2.2	2.3	4.9
	95th %ile CV	**3.1**	**4.2**	6.8	13

Daily relative pair difference (%) was calculated from the absolute difference of the two daily measurements divided by the mean of the two measurements times 100.

Campaign CV (%) was the mean of 6 independent preparations in 3 runs (duplicates per run)