Journal of Radioanalytical and Nuclear Chemistry (2022) 331:4929–4938
https://doi.org/10.1007/s10967-022-08618-1

Who's who among radioactive analysts statistical and graphical analysis of variance in measurement accuracy across radioanalytical laboratories

Kenneth G. W. Inn[1,2] · **Nateé Johnson**[3] · **Iisa Outola**[2,5] · **James Filliben**[4]

Received: 28 April 2022 / Accepted: 1 October 2022 / Published online: 25 October 2022

Abstract

The National Institute of Standards and Technology Radiochemistry Intercomparison Program [NRIP] has been assessing radioanalytical laboratory capabilities. This study evaluates performance with respect to 15 radionuclides, 24 laboratories, 5 matrices, and measurement "outliers" over a span of 10 years. Results indicated: (1) the data is best characterized using medians and treating "outliers" as part of non-normal distributions; (2) results among the factors of matrices and over 10 years were in good control, while the factors of radionuclides and laboratory showed larger variation in performance; (3) of all the radionuclides tested alpha-emitter analysis were in good control while beta-emitting ^{90}Sr analysis showed the most variance; (4) laboratories that tended to be in control continued to demonstrate consistent performance while some lagging laboratories improved their performance after 2–3 years of experience in the program; (5) spiked urine and glass fiber filters showed the best results while spiked synthetic fecal and soil sample results were more problematic; and (6) there was little variation of unweighted median laboratory performance over the 10 years of testing. These results provides NRIP and the participating laboratories with much needed feedback to address problem areas that were uncovered, share advanced radioanalytical methods to improve lagging laboratory performance, and provide a platform for discussion of new radionuclides, matrices, interferences, and approaches for the pass/fail criteria and uncertainty reporting requirements to improve and expand the performance evaluation program.

Keywords Measurement traceability · Performance testing · Radionuclides · Matrices · Laboratories · Time

Introduction

There are many applications for radionuclide measurements. In a number of instances long-term monitoring programs require many thousands of measurements for safe use of radioactivity and nuclear technology, and there are instances where only a few measurements are made for critical national security decisions. In all of these cases, measurement capabilities must be of the highest integrity (accuracry and precision) and reliability (reproducible). Independent performance testing, and more importantly, independent traceability testing provides the link to accuracy to the SI unit of radioactivity—the Becquerel [Bq].

The National Institute of Standards and Technology [NIST] has been conducting the NIST Radiochemistry Intercomparison Program [NRIP] traceability evaluation program where the blind testing results provides the participating laboratories with the information to determine if their measurement processes are in long-term statistical control, improve capabilities and verify their quality control processes to meet programmatic measurement quality objectives, and to assess their capabilities to respond to time-sensitive emergency situations.

✉ Kenneth G. W. Inn
kgenwahinn@gmail.com

1 K&E Inn Ovations, Inc., 91-1329 Kuano'o Street, Ewa Beach, HI 96706, USA

2 Formerly National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

3 Formerly University of Maryland, Baltimore County, 1000 Hilltop Cir, Baltimore, MD 21250, USA

4 National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

5 Currently Radiation and Nuclear Safety Authority (STUK), Laippatie 4, 00880 Helsinki, Finland

A broad suite of fifteen radionuclides for gross radioactivity, alpha, beta, and gamma assay are quantitatively spiked into five replicate acidified water, synthetic urine, synthetic feces, glass fiber air filters, and soil samples for the participant laboratories' measurements. The attached Supplemental Information is the NRIP Statement of Work [SOW] for 2006. This SOW details the potential certified nuclides that could be use to spike the test samples, the maximum Bq/sample for each radionuclide for both the 60 day turnaround and 8 h turnaround programs, and traceability testing criteria. These NRIP program requirements were established during detailed discussions between the participating laboratories and NIST. Considerations for the agreed upon activity levels to be spiked into the test samples included the participating laboratories' measurement sensitivity, potential for laboratory contamination, and sufficient activity levels to make sufficiently precise measurements to meet traceability testing criteria.

Since its beginning in 1997 more than twenty laboratories have participated in NRIP, and the program continues today. The routine monitoring aspect of the program has a measurement turn around time of two months while it is limited to eight hours for the emergency response measurements.

While providing radioactivity traceability is a unique mission to NIST, it does not provide a regulatory pass/fail assessment criteria for traceability. NIST depends on the individual programs under which the participating laboratories are operating to provide the necessary intra-program specific assessment criteria. In the absence of programmatically declared assessment criteria, NIST invokes with agreement by the participating laboratories, the use of the consensus ANSI N42.22 & ANSI N13.30 testing criteria [1, 2].

Periodically, NIST will conduct a global evaluation of the results received by NRIP [3–5]. This paper summarizes the 10 year evaluation for the routine monitoring part of the program. The laboratory performance issues to be evaluated were the difference [”bias”] from the NIST certified value, variation, and relative importance of each of the following factors:

- Which matrix demonstrated best agreement with NIST certified values? Worst?
- Which year demonstrated best agreement with NIST certified values? Worst?
- Which radionuclide demonstrated best agreement with NIST certified values? Worst?
- Were there any outliers affecting data?
- Which laboratory demonstrated best agreement with NIST certified values? Worst?

Methods

The data from the routine monitoring section of NRIP was selected for assessment because it is less affected by the effects of counting uncertainty components that could dominate measurement uncertainty under emergency response time frame. All of the 1269 measurement results collected from the 24 participating laboratories across the 10 years [1997–2006] for all 5 matrices and 15 radionuclides were compiled into an Access database. No “outliers” were initially excluded to assess their effects on the data distributions. The Access database was then exported to Dataplot [6] for statistical evaluation where the variations of the mean, median, and data distributions could be visualized.

Results and discussion

The statistical evaluation of measurement results collected by NRIP over ten years are summarized in Figs. 1, 2, 3, 4 and 5.

What was the relative importance of each NIST programmatic factor?

Figure 1 depicts the variation of the relative mean difference from the NIST certified values as a function of laboratory, year, matrix and radionuclide. The factors that showed the largest mean deviation from the NIST values were due to radionuclides and laboratories [up to 50%] while year and matrices showed little effect [maximum < 5%].

Which matrix showed the best agreement with NIST certified values? Worst?

Figure 2a-d present the results sorted by test matrices. The data is displayed from left to right for acidified water [AW], glass fiber [GF] air filters, synthetic fecal material [SF], spiked soil [SS], and synthetic urine [SU]. Figure 2a shows that the bulk of the 1269 measurement result are within ~ 50% of the NIST values across all matrices but there are deviations as large as ~ 340%, predominantly due to ^{90}Sr measurements. Figure 2b and c shows that the unweighted mean differences among the matrices from which measurement data were non-normally distributed, resulting in the median depiction being preferred. The maximum median deviation among the matrices was ~ 3.5%. Examination of Fig. 2c, indicated that the best unweighted median results come from analyses of radionuclides on glass fiber air filters, acidified water and synthetic urine while synthetic fecal and spiked soil matrices were more problematic.

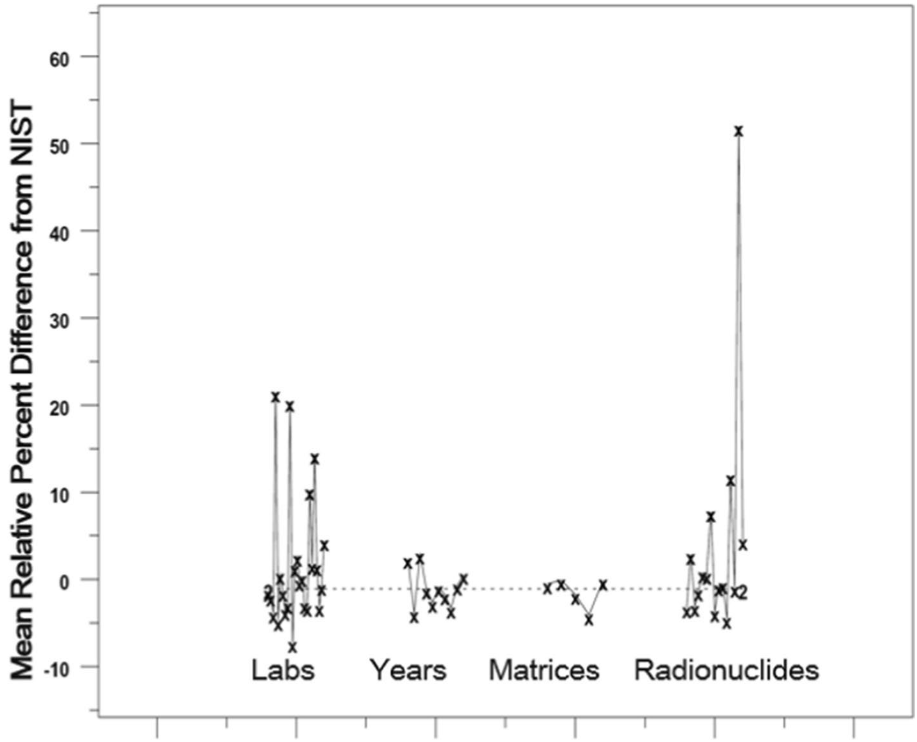


Fig. 1 What is the relative importance of each factor? The mean relative measurement difference from the certified NIST values [Y-axis] for the 24 participating laboratories, the years from 1997 to 2006, the test 5 matrices and the 24 radionuclides [X-axis] are presented

Non-normal measurement result distributions were identified using Normal Probability Plots like that depicted in Fig. 2d where the 5 points on the right side of the curve deviate significantly from a normal distribution. The non-normal depiction of the measurement data distribution supports the preference for the use of the median characterization of the data. Identification of these 5 "outliers" provides an opportunity for future "root-cause" study and analytical method adjustments when warranted.

Which year showed the best agreement with NIST certified values? Worst?

Figure 3a and b show the variation of the relative unweighted median and mean difference from the NIST certified values. The difference between Fig. 3a and b indicates that there is a difference between the mean vs. median methods of doing the evaluation. This in turn points out that the distribution of yearly unweighted means is non-normal and that the median is preferred. For Fig. 3a, the maximum average deviation of median values from the NIST certified values was only ~ 3%, thereby indicating that time is not a major factor.

Which radionuclide showed the best agreement with NIST certified values? Worst?

Fi01gure 4a represents the results among the fifteen certified radionuclides [^{133}Ba, ^{137}Cs, ^{152}Eu, ^{230}Th, ^{232}Th, ^{234}U, ^{235}U, ^{238}Pu, ^{238}U, ^{239}Pu + ^{240}Pu, ^{241}Am, ^{54}Mn, ^{60}Co, ^{65}Zn, ^{90}Sr] that were gravimetrically spiked into all five NRIP matrices and verified by radiochemical assay. Globally, across all matrices the most challenging radioanalysis for the participating laboratories was for ^{90}Sr. Even for acidified water samples, ^{90}Sr measurements were challenging and had a wider dispersion of results than for the other radionuclides as shown in Fig. 4b. While radiochemical measurements of ^{238}U, ^{234}U, and ^{239}Pu + ^{240}Pu has the advantages of the use of internal tracers and alpha energy spectroscopic based measurement systems, the measurement of ^{90}Sr is complicated by its complex radiochemistry and continuous energy spectrum of the beta particles that necessitates the use of non-spectroscopic detector measurement/calibration processes. Radionuclides determined by gamma spectrometry were fairly good, but care needed to be taken to calibrate the instruments adequately for geometry, sample density, and corrections for true coincidence summing.

A note is given here that gross alpha/beta results are not included in the Figures and will not be discussed in detail. However, as a general comment, the reported results varied

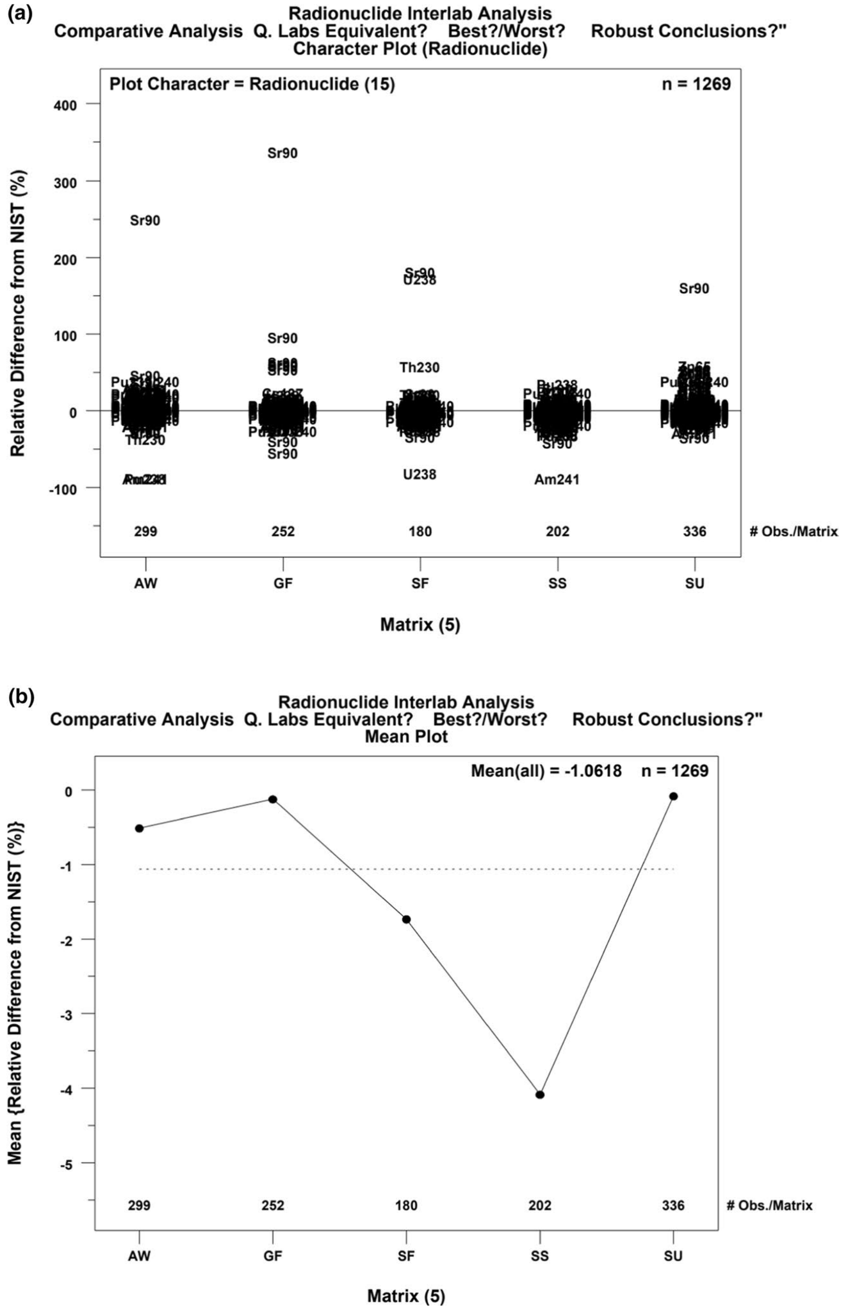


Fig. 2 Which matrix demonstrated best agreement with NIST certified values? Worst?. **a** The effect of test matrix on variance performance. Y-axis = Relative difference for all 1269 measurements from the NIST certified value for each matrix. X-axis = matrix where AW = Acidified Water; GF = Glass Fiber Air Filter; SF = Spiked Fecal; SS = Spiked Soil; SU = Spiked Urine. The solid horizontal line is the NIST certified value. Also noted are the number of measurement observations for each matrix. **b** The effect of test matrix on the mean measurement performance. Y-axis = Mean relative difference for all 1269 measurements from the NIST certified value for each matrix. X-axis = Matrix where AW = Acidified Water; GF = Glass Fiber Air Filter; SF = Spiked Fecal; SS = Spiked Soil; SU = Spiked Urine. The dashed horizontal is the mean -1.06% measurement difference from the NIST certified value across all matrices for all measurements. Also noted are the number of measurement observations for each matrix. **c** The effect of test matrix on the median measurement performance. Y-axis = Relative difference from the NIST certified value for each matrix. X-axis = Matrix where AW = Acidified Water; GF = Glass Fiber Air Filter; SF = Spiked Fecal; SS = Spiked Soil; SU = Spiked Urine. The dashed horizontal is the median − 1.8% measurement difference from the NIST certified value across all matrices for all 1269 measurements. Also noted are the number of measurement observations for each matrix. **d** Outlier Identification using Normal Probability Plots. Y-axis = ranked % measurement difference from the NIST certified value for all measurements. X-axis = Standard Deviation

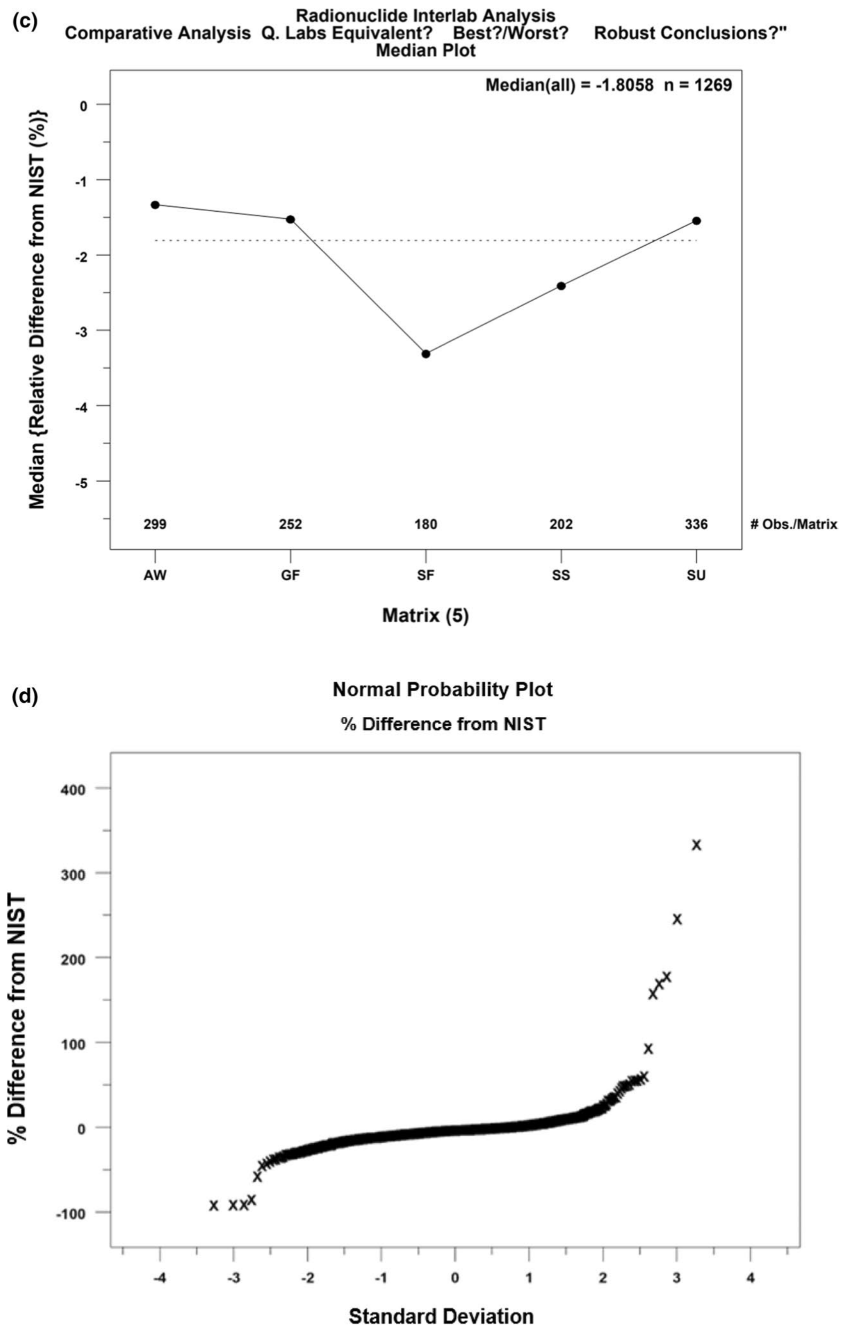


Fig. 2 (continued)

widely from the NIST certified values [> 100%]. It is suspected that the discrepancy is mainly due to the difference in the average energy from the fifteen radionuclides in the NRIP test cocktail from that used for the participating laboratories' instrument calibration. So as a general caution for the use of gross alpha/beta field screening measurements, the

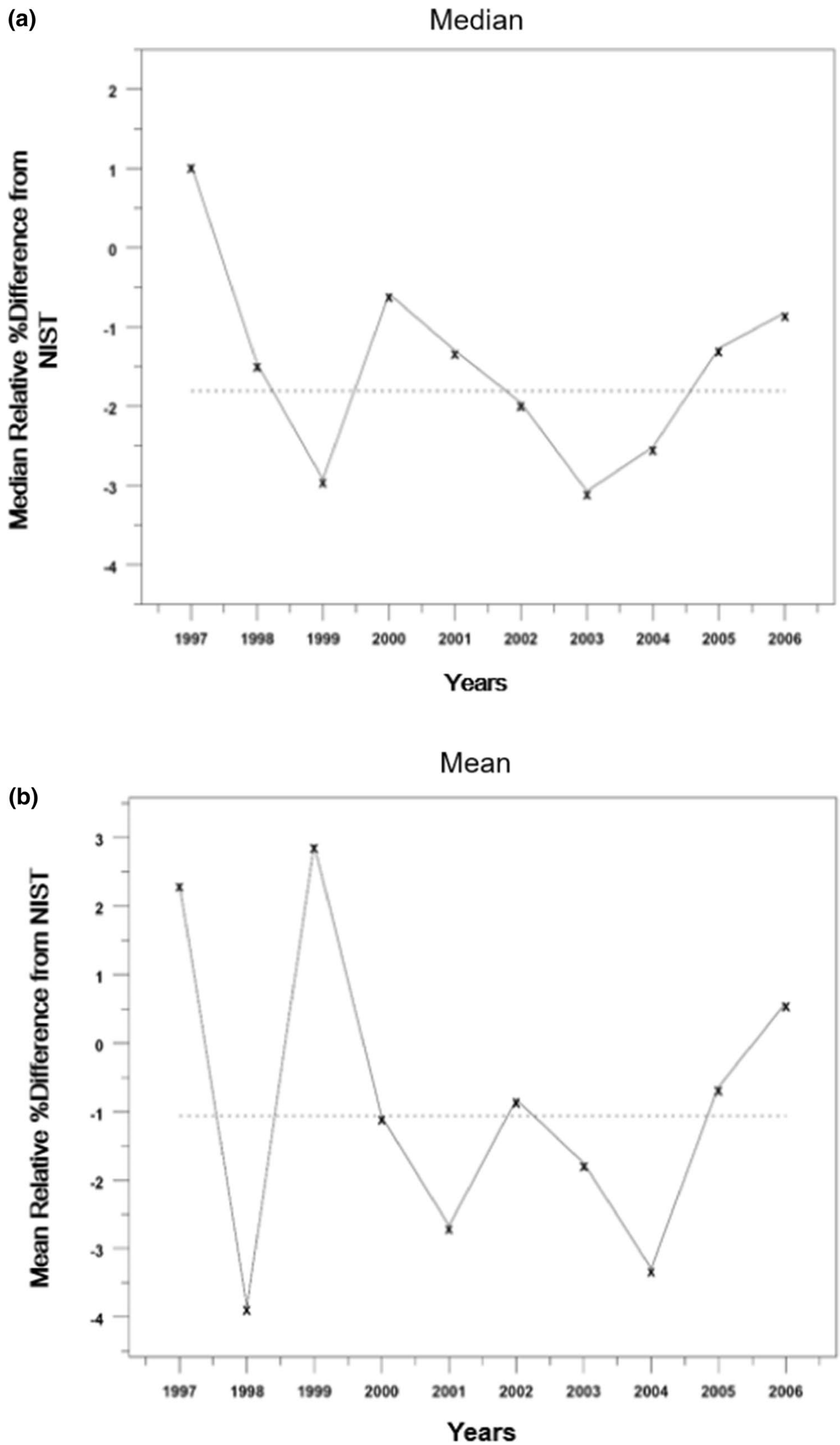


Fig. 3 **a** Which Year demonstrated best agreement with NIST certified values? Worst? The effect of year on the median measurement performance. Y-axis = Median relative measurement difference for all measurements from the NIST certified value, and X-axis = each of the 10 years. The dashed horizontal line is the median measurement difference from the NIST certified value across all years. **a** Which Year demonstrated best agreement with NIST certified values? Worst? The effect of year on the mean measurement performance. Y-axis = Mean relative measurement difference for all measurements from the NIST certified value, and X-axis = each of the 10 years. The dashed horizontal line is the mean measurement difference from the NIST certified value across all years

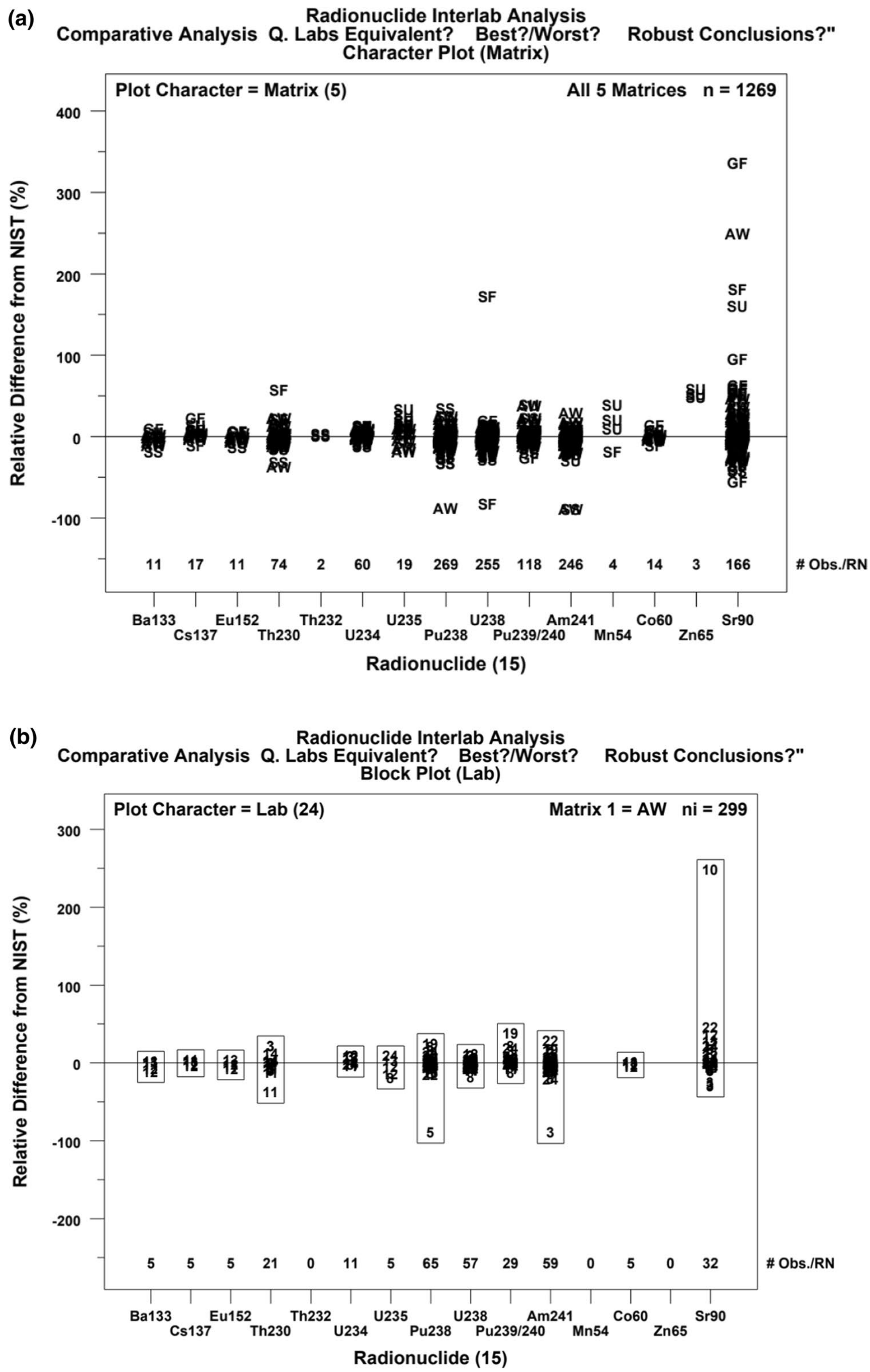


Fig. 4 **a** What Nuclide demonstrated best agreement with NIST certified values? Worst? The effect of test radionuclide on measurement performance. Y-axis = Relative difference of 1269 measurement results for all 5 matrices from the NIST certified values, and X-axis = all 15 radionuclides. The solid horizontal line is 0 percent difference from the NIST certified value. Also noted are the number of measurement observations for each radionuclide. **b** The effect of test radionuclide on measurement performance for the acidified water matrix. Y-axis = Relative difference of 299 measurement results from the NIST certified values, and X-axis = all 15 test radionuclides. The solid horizontal line is 0 percent difference from the NIST certified value. Also noted are the number of measurement observations for each radionuclide

accuracy can be greatly improved by calibrating the measurement instruments with sources of appropriate energy, or by applying an appropriate correction factor to account for the difference in calibration energy from real samples.

Were there any outliers affecting data?

As pointed out previously, "outliers" represented a very small fraction of all of the measurement results reported

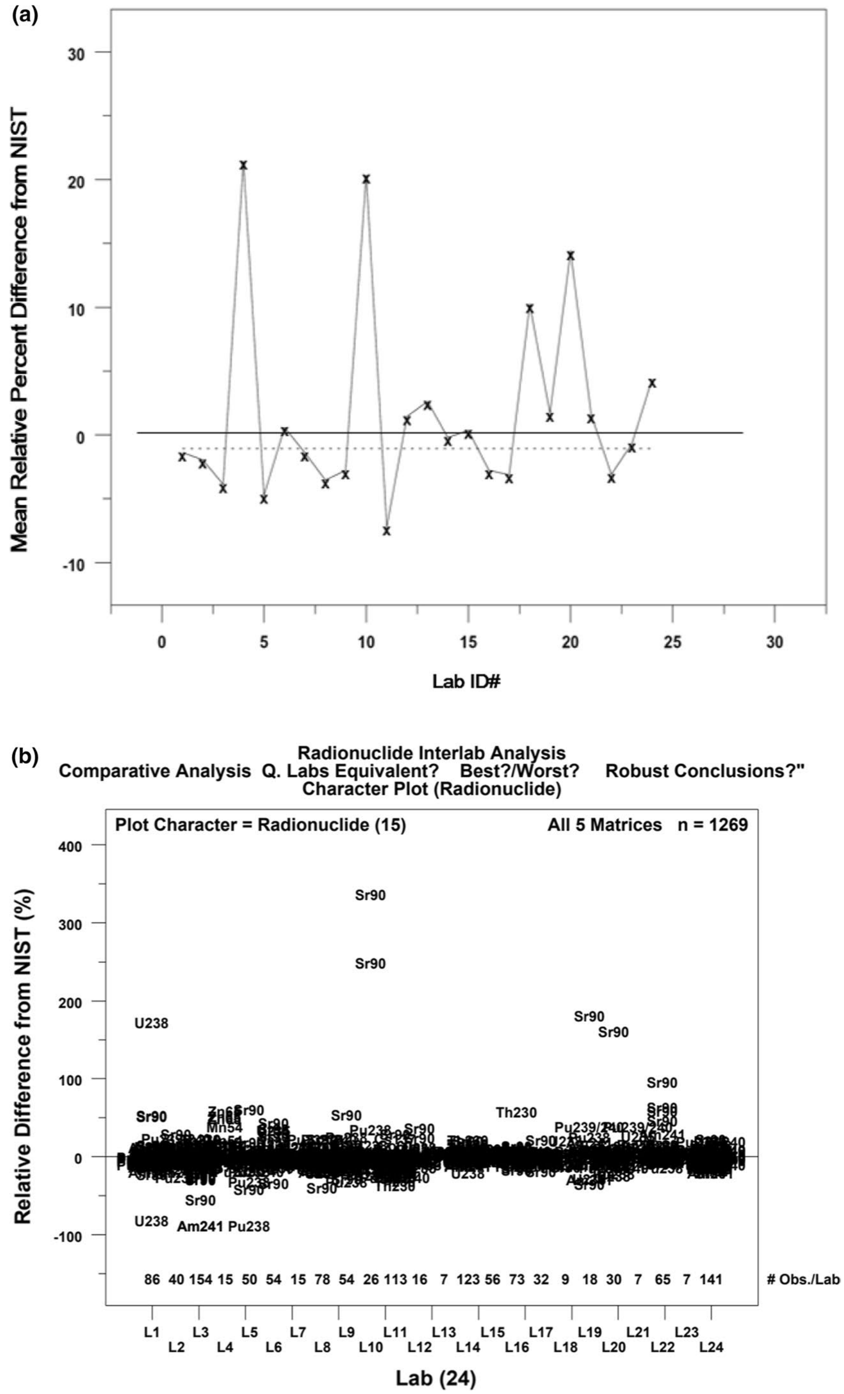
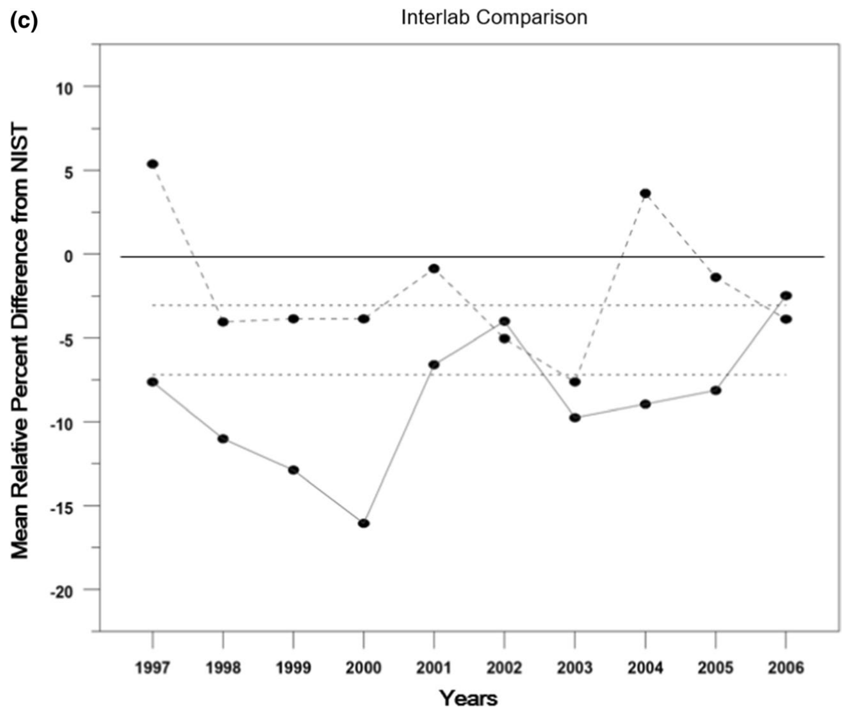


Fig. 5 **a** Mean Performance of the 25 participating laboatories. Y-axis = Mean relative difference from the NIST certified value [Y-axis] for all measurements, and X-axis = each of 24 participating laboratories. The solid horizontal line is 0 percent difference from the NIST certified value, and the dashed line represents the mean difference for the combined laboratory results. **b** Laboratory performance for individual radionuclide measurement. Y-axis = Relative difference of the 1269 measurements from the NIST certified value for all 15 radionuclides, and X-axis = each of 24 participating laboratories. The solid horizontal line is 0 percent difference from the NIST certified value. Also noted are the number of measurement observations by each laboratory. **c** How Do Long-Term Participants Compare? The solid line between dots reflect the relative mean measurement difference from the NIST certified value [Y-axis] across all radionuclides and matrices for lab 11 and the dashed line between dots represents the mean results for lab 22 from 1997 to 2006 [X-axis]. The upper horizontal dashed line is the mean results over the 1977–2006 timeframe for lab 22 and the lower horizontal dashed line is for lab 11's performance over the same timeframe. The solid horizontal line at 0% difference represent the NIST certified values

Fig. 5 (continued)



during this ten year study. However, the topic of discovering underlying bias will need further examination in the future.

As it turns out, in general, this study shows that only an occassional"outlier" was noted to the point where it is within statistical expectations. 9 out of 1269 measurements were visually assessed as potential"outliers." This is a rate of ~ 0.7%. An even better picture would be the 5"outliers" identified from the Normal Probability Plot on Fig. 2d. This assesssment indicates an"outlier" rate of 0.4%. It could be argued that the rate of"outliers" seen in this study was confirmation of good measurement control among the participating laboratories.

Which laboratory showed the best agreement with NIST certified values? Worst?

Figure 5a presents the mean laboratory performance over time, matrices and radionuclides. Over all the participating laboratories the combined average performance was within ~ 2% of the NIST certified values. However, there are two laboratories with performances beyond 10% and two more in excess of 20% of the NIST values. The other 20 laboratories are well within 5% of the certified values.

Figure 5b present the laboratory's radionuclide measurement performance across all matrices and years. While it can be seen that the bulk of the results hover around "0" deviation from the certified values, there are ~ 9 apparent "outlier" entries, mostly due to ^{90}Sr. Root-cause investigations of these "outliers" could benefit the participant laboratories to improve their process control, reduce analytical sources of dark bias, and reduce their measurement uncertainties.

Figure 5c shows the historical performance for laboratories 11 and 22. Laboratory 22 demonstrates consistent results with an overall deviation of ~ 3% over 10 years. By contrast, laboratory 11 started out with its first four years of evaluation with deviations from the certified values on the order of 12% but improved its performance over the next 6 years to show capabilities approaching that of laboratory 22. Participation in the NRIP traceability evaluation program provides documentation of continued high quality performance and also of improving performance.

Conclusions

The major conclusions from this assessment were: [1] no significant change in global data over time, [2] there were only a small number of "outlier" measurement results over all the factors investigated [3] global areas for improvement were identified, and [4] NRIP was important for laboratories to: [a] track performance, [b] improve performance, and [c] enhance quality assurance/control.

Further research identified by this assessment were: [1] more granular inspection of the database on a laboratory specific basis, e.g., dark sources of analytical bias, [2] root-cause basis for method improvements, [3] attention to answer "How good is good enough?" to establish programmatic-based traceability pass/fail criteria, and [4] Emergency Response, e.g., development of faster radiochemical procedures and more accurate measurement methods so more time can be spent on counting to reduce it's contribution to combined uncertainty, and [5] advanced measurement methods for higher precision and accuracy. Additionally, the results from this study provides a platform for discussion of new radionuclides, matrices, and interferences to be incuded in future NRIP exercises.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10967-022-08618-1.

Acknowledgements This publication is dedicated to the memory of Dr. James John Filliben, Sr. for his crucial statistical insights, contributions and encouragement towards this work and for pass collaborations. The authors gratefully acknowledge the technical contributions, insightful and patient comments and suggestions from Jerome LaRosa, Svetlana Nour, Matthew Mille, and Loleta Dixon. The authors also gratefully acknowledge the generous support provided by Marc Desrosiers, the NIST Summer Undergraduate Research Fellowship (SURF), the National Institute of Standards and Technology, the National Science Foundation (NSF), and University of Maryland, Baltimore County (UMBC).

References

1. American National Standard Institute (ANSI) (1995) American national standard—traceability of radioactive sources to the national institute of standards and technology (NIST) and associated instrument quality control, ANSI N42.22–1995, institute of electrical and electronics engineers (IEEE), New York, NY.
2. American National Standard Institute (ANSI) (1996) American national standard—performance criteria for radiobioassay, ANSI N13.30–1996, health physics society, McLean, VA
3. Wu Z, Inn KGW, McMahon CA, Karam LR (2002) NIST radiochemistry intercomparison program: a summary of four-year performance evaluation study. Appl Radiat Isot 56:379–385
4. Inn KGW, Outola I, Nour S, Kurosaki H, Albin L, Berne A (2006) Radioanalytical emergency response exercise. J Radioanal Nuclear Chem 269(2):351–360
5. Nour S, Inn KGW, LaRosa J (2011) Emergency radiobioassay preparedness exercises through the NIST radiochemistry intercomparison program. Health Phys 101(2):170–175
6. Heckert A, Filliben J, (2003) Dataplot Reference Manual, NIST Handbook, 148.