Multi-centre evaluation of accuracy and reproducibility of planar and SPECT image quantification: An IAEA phantom study

Brian E. Zimmerman[1,*], Darko Grošev[2], Irène Buvat[3], Marco A. Coca Pérez[4], Eric C. Frey[5], Alan Green[6], Anchali Krisanachinda[7], Michael Lassmann[8], Michael Ljungberg[9], Lorena Pozzo[10], Kamila Afroj Quadir[11], Mariella A. Terán Gretter[12], Johann Van Staden[13], Gian Luca Poli[14]

[1] National Institute of Standards and Technology, Gaithersburg, MD, USA
[2] University Hospital Centre Zagreb, Zagreb, Croatia
[3] Service Hospitalier Frédéric Joliot, Paris, France
[4] Medscan Concepción, Concepción, Chile
[5] Johns Hopkins University, Baltimore, USA
[6] National Physical Laboratory, Teddington, UK
[7] Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand
[8] University Würzburg, Würzburg, Germany
[9] Lund University, Lund, Sweden
[10] Institute of Energy and Nuclear Research, São Paulo, Brazil
[11] National Institute of Nuclear Medicine & Allied Sciences, Dhaka, Bangladesh
[12] Universidad de la República, Montevideo, Uruguay
[13] University of the Free State, Bloemfontein, South Africa
[14] International Atomic Energy Agency, Vienna, Austria

Received 4 December 2015; accepted 30 March 2016

Abstract

Accurate quantitation of activity provides the basis for internal dosimetry of targeted radionuclide therapies. This study investigated quantitative imaging capabilities at sites with a variety of experience and equipment and assessed levels of errors in activity quantitation in Single-Photon Emission Computed Tomography (SPECT) and planar imaging. Participants from 9 countries took part in a comparison in which planar, SPECT and SPECT with X ray computed tomography (SPECT-CT) imaging were used to quantify activities of four epoxy-filled cylinders containing ^{133}Ba, which was chosen as a surrogate for ^{131}I. The sources, with nominal volumes of 2, 4, 6 and 23 mL, were calibrated for ^{133}Ba activity by the National Institute of Standards and Technology, but the activity was initially unknown to the participants. Imaging was performed in a cylindrical phantom filled with water. Two trials were

Multizentrische Evaluierung der Genauigkeit und Präzision bei der Quantifizierung planarer Bildgebung und SPECT: Eine Phantomstudie der IAEA

Zusammenfassung

Eine genaue Quantifizierung der Aktivität ist die Grundlage für die interne Dosimetrie bei der Radiotherapie. Diese Studie beschreibt die Fähigkeiten zur quantitativen Bildgebung in Zentren unterschiedlicher Erfahrung und mit unterschiedlicher Ausstattung. Das Ziel war die Abschätzung der Fehler bei der Aktivitätsquantifizierung mittels SPECT und planarer Bildgebung. An dieser Vergleichsstudie nahmen Zentren aus 9 Ländern teil. Es wurden planare Bildgebung, SPECT und SPECT/CT verwendet, um Aktivitäten in epoxidharzgefüllten Zylindern,

* Corresponding author: Brian E. Zimmerman, National Institute of Standards and Technology, Physical Measurement Laboratory, 100 Bureau Drive, Stop 8462, Gaithersburg, MD 20899-8462, USA. Tel.: +1 301 975 4338.
E-mail: bez@nist.gov (B.E. Zimmerman).

carried out in which the participants first estimated the activities using their local standard protocols, and then repeated the measurements using a standardized acquisition and analysis protocol. Finally, processing of the imaging data from the second trial was repeated by a single centre using a fixed protocol. In the first trial, the activities were underestimated by about 15% with planar imaging. SPECT with Chang's first order attenuation correction (Chang-AC) and SPECT-CT overestimated the activity by about 10%. The second trial showed moderate improvements in accuracy and variability. Planar imaging was subject to methodological errors, e.g., in the use of a transmission scan for attenuation correction. The use of Chang-AC was subject to variability from the definition of phantom contours. The project demonstrated the need for training and standardized protocols to achieve good levels of quantitative accuracy and precision in a multicentre setting. Absolute quantification of simple objects with no background was possible with the strictest protocol to about 6% with planar imaging and SPECT (with Chang-AC) and within 2% for SPECT-CT.

Keywords: Comparison, quantitative imaging, SPECT, planar

die ^{133}Ba als Surrogat für ^{131}I enthielten, zu quantifizieren. Die Quellen mit nominalen Volumina von 2, 4, 6 und 23 mL wurden vom National Institute of Standards und Technology (NIST) für ^{133}Ba kalibriert. Die Aktivität der Quellen war den Teilnehmern zunächst unbekannt. Die Bildgebung erfolgte in einem zylindrischen mit Wasser gefüllten Phantom. Es wurden zwei Studien durchgeführt, bei denen die Teilnehmer zunächst die Aktivitäten gemäß ihren lokalen Standardprotokollen bestimmen sollten. Danach wurde die Messung mit einem standardisierten Akquisitions- und Auswerteprotokoll durchgeführt. Die Auswertung der zweiten Studie wurde zusätzlich zentral mit einem genau festgelegten Protokoll wiederholt.

In der ersten Studie wurden die Aktivitäten mit planarer Bildgebung um etwa 15% unterschätzt. SPECT mit einer Schwächungskorrektur nach Chang und SPECT-CT überschätzten die Aktivität um ca. 10%. Die zweite Studie zeigte moderate Verbesserungen in der Genauigkeit und Variabilität. Planare Bildgebung unterlag methodischen Fehlern, z.B. bei der Verwendung eines Transmissionsscans zur Schwächungskorrektur. Die Verwendung der Schwächungskorrektur nach Chang führte zu Ergebnissen aufgrund der Schwierigkeit, die Grenzen des Phantoms genau zu definieren.

Das Projekt zeigte die Notwendigkeit von Schulungen und standardisierten Protokollen in einer multizentrischen Studie, um gute quantitative Genauigkeit und Präzision zu erzielen. Absolute Quantifizierung von einfachen Objekten ohne Hintergrund war, mit einem genau festgelegten Protokoll, mit einer Genauigkeit von etwa 6% mittels planarer Bildgebung und SPECT (mit Schwächungskorrektur nach Chang) und weniger als 2% für SPECT-CT möglich.

Schlüsselwörter: Vergleichsstudie, quantitative Bildgebung, SPECT, planar

1 Introduction

Nuclear medicine imaging has the potential to provide quantitative information about the distribution of an injected radiopharmaceutical in the body. Successive scans on the same patient at different times can provide data on the radiopharmaceutical biokinetics, i.e., the change in the distribution of the radiopharmaceutical in the body over time [1]. There are several nuclear medicine applications that require images that are quantitatively accurate in an absolute sense [2–5]. For example, absolute quantification is needed in the therapeutic use of radiopharmaceuticals, where the purpose is to evaluate the dose-effect relationship [6–9].

Accurate quantitation of the uptake of injected activity in various organs over time provides the basis for internal dosimetry. This is especially important to rigorously optimize the therapeutic use of radiopharmaceuticals, where patient specific dosimetry is often a legal obligation due to the high absorbed doses [10]. As a consequence, there is a need for harmonized protocols or guidelines for acquiring quantitative information from nuclear medicine imaging procedures.

Over the past two decades, there has been much research and progress in developing methods for accurately quantifying nuclear medicine images obtained from scintillation cameras [11,12]. However, adoption of these methods by clinics has been slow. Generating images from scintillation cameras suitable for quantitative tasks requires additional attention to data acquisition and processing compared to those used solely for qualitative (i.e., visual) interpretation. Moreover, absolute quantitation is more challenging than relative quantitation as it imposes greater demands on the accuracy of corrections for scatter, attenuation, partial volume, and other effects. In

addition, there is the need for calibration of the imaging system and activity measurement instruments [13].

Achieving absolute quantitation requires appropriate equipment, software, and human resources. The level of these requirements depends on the imaging task. For example, quantifying activity in a tumour in the lungs requires more sophisticated resources than quantifying whole body activity. However, the levels of resources needed for each type of imaging procedure have not been systematically investigated and are therefore not widely available.

Guidance documents and training are needed to take full advantage of the capabilities of nuclear medicine instruments and to understand their limitations, which can ultimately lead to better quantitative practice worldwide. The Medical Internal Radiation Dosimetry Committee of the Society of Nuclear Medicine and Molecular Imaging has recently published a document providing guidance for Single Photon Emission Computed Tomography (SPECT) quantification and more detailed recommendations for ^{131}I [14,15]. To further address this need, the International Atomic Energy Agency (IAEA) initiated the Coordinated Research Project E2.10.07, entitled "Development of Quantitative Nuclear Medicine Imaging for Patient Specific Dosimetry". A report on nuclear medicine imaging quantification [16] was the first output of this project, while the present study comprised the majority of work undertaken by the participants.

The purposes of the study described herein were to: (i) investigate the capabilities for quantitative single photon imaging at different sites worldwide, including sites where resources are limited; (ii) develop and test quantitative imaging methods in nuclear medicine practice; (iii) assess the need for standardization and harmonization of quantitative nuclear medicine on an international scale; and (iv) assess the achievable accuracy of absolute activity quantitation for different nuclear medicine methodologies in a range of sites with various levels of available resources.

This paper describes the results of an image quantification comparison study performed between 9 groups of medical physicists working in institutions in different countries. The participants at the various sites each imaged a phantom containing a different set of ^{133}Ba source inserts calibrated by the U.S. National Institute of Standards and Technology (NIST). Images were obtained and activities in the sources were quantified using three different modalities: planar, SPECT and SPECT with X-ray computed tomography (CT).

2 Methods

2.1 Sources

The primary goal of the study was to assess the accuracy and variability of absolute activity estimates obtained via scintillation camera imaging at multiple international sites. Thus, multiple sets of accurately calibrated sources were needed. Iodine-131 is widely used for radiopharmaceutical therapy, and it was decided to base the comparison on this radionuclide. However, because of the relatively short half-life ($T_{1/2}$ = 8.0233(19) d) [17] of ^{131}I and the large distances between the participating laboratories, it was deemed impractical to prepare, calibrate, and ship phantoms using ^{131}I. Moreover, solid epoxy sources were required for the comparison to avoid logistical problems associated with shipping liquid radioactive sources internationally.

Barium-133 was chosen as the surrogate because of its long half-life of 10.540(6) years [17] and similarities between the ^{133}Ba and ^{131}I decay schemes. The most abundant γ-ray in the decay of ^{133}Ba at 356 keV is similar in energy and emission probability to the most abundant γ-ray in the decay of ^{131}I at 364 keV. The primary differences in the decay schemes are the presence of 637 keV and 722 keV γ-rays in ^{131}I decay that are not found in ^{133}Ba, as well as the possible production of bremsstrahlung from the β^- electrons emitted in the decay of ^{131}I. These high-energy photons can contribute to the main photopeak energy window from Compton scatter events (in the patient or in the collimator), as well as events undergoing both collimator septal penetration and Compton scattering [18]. Also, ^{133}Ba has photons with energy 303 keV that need to be taken into account when selecting energy windows for scatter compensation.

The phantom sources that were distributed to the participants were the same as those described in [19]. Briefly, each distributed set consisted of four ^{133}Ba epoxy-filled poly(methyl methacrylate) (PMMA) cylinders, each having an active epoxy length of 3.8 cm and inside diameters of 0.794 cm, 1.27 cm, 1.43 cm, and 2.86 cm to give nominal volumes of 2 mL, 4 mL, 6 mL, 23 mL, respectively. These were given the series designations A, B, C, and D, respectively. The dimensions of the sources were chosen to evaluate quantification in the context of different levels of partial volume effect. The diameter of the largest source was chosen to be sufficiently large so as to be essentially free of such effects for most modern SPECT cameras.

For the A, B, and C series, the calibrated activity concentration was (0.199 ± 0.003) MBq g^{-1} at the established reference time. For safety and transport reasons, the activity concentration of the D series sources was chosen to be a factor of four lower, and thus the calibrated activity concentration at the same reference time was (0.050 ± 0.001) MBq g^{-1}. The relative standard uncertainties on the activities for the A, B and C series were 1.43%, while those for the D series were somewhat higher at 1.74%. For each individual source, the total contained activity was calculated by multiplying the appropriate activity concentration by the respective measured epoxy mass, which was determined with an uncertainty of about 0.2%.

A set of four sources (one from each series) was dispatched to each participant between May 2011 and February 2013. Measurements were made at the convenience of the participants, however, it was expected that results would be submitted no more than 2 months after receipt of the sources.

Table 1
Scintillation cameras used in the multicentre comparison.

Participant	Manufacturer	Model	Year of installation
Bangladesh	Siemens	ECAM	2006
Brazil	Siemens	Symbia T16 SPECT-CT	2010
	General Electric	Infinia Hawkeye 4	2007
Croatia	Siemens	Symbia E/Symbia T2 SPECT-CT	2010
Cuba	Mediso	Nucline Spirit DH-V	2009
Germany	Siemens	Symbia T2 SPECT-CT[a]	2008
South Africa	Siemens	Symbia T SPECT-CT	2007
Thailand	Siemens	Symbia T6 SPECT-CT	2012
USA	Siemens	Symbia T16 SPECT-CT	2011
Uruguay	Mediso	Nucline Spirit-DHV	2007

[a] Crystal thickness 15.9 mm.

2.2 Participants and equipment

The different institutions contributing to the measured data were from Bangladesh, Brazil, Croatia, Cuba, Germany, South Africa, Thailand, United States of America, and Uruguay. All participants responsible for the measurement were trained medical physicists with experience in quantitative imaging in their local departments. The choice of institutes was based on evaluations by the IAEA of applications received to participate in the larger Coordinated Research Project (discussed above), with the exception of participants from the United States and Germany, which were recruited based on their extensive expertise in this area. Because this comparison study formed a relatively small part of the overall project, no specific effort was made to achieve diversity in terms of scanner manufacturer, model or type.

Participation in this study required access to a scintillation camera system with planar and SPECT imaging capabilities and methods for attenuation and scatter correction. Since ^{133}Ba was used as a substitute for ^{131}I, a high-energy collimator was a prerequisite for the study. The oldest camera was installed in 2006 and five systems were equipped with CT acquisition capabilities and CT-based attenuation correction. The scintillation cameras were manufactured by three vendors (Siemens, General Electric, and Mediso) and were all equipped with 9.5 mm crystals, except one, which had a 15.9 mm thick detector. All the systems were subject to quality assurance programmes according to local requirements. The project also required the ability to acquire images in multiple energy windows to apply the Triple Energy Window [20] method for scatter correction. Details on the scanners used in the study are reported in Table 1. The sensitivity factors (i.e., calibration coefficients), were measured by the participants using their respective B-series source for both planar and SPECT. For each participant, the measured sensitivities for the two different techniques were generally consistent.

Activities were calculated from the observed count rates and the sensitivity factors using well-established methods [16].

2.3 Acquisition and image analysis protocols

For this study two trials were initiated. For the first trial, participants were expected to develop their own protocol to quantify the activities in the four supplied ^{133}Ba sources placed in the phantom by planar and SPECT methods. No acquisition or processing parameters were specified. The only specific training provided was a draft of the IAEA guidance document described above [16]. The activities of the sources were not provided to the participants. For calibration, participants were instructed to use their B-series sources to calibrate their systems using the activity value measured using their local activity calibrator. The choice of source was a compromise between partial volume effects and the introduction of additional uncertainty due to attenuation and scattering. Reconstructed SPECT counts were used to determine the sensitivity factor. The time between calibration and measurements was variable and was left to the discretion of each participating institute.

The ^{133}Ba sources were designed to be used in conjunction with a standard 22 cm diameter cylindrical phantom, one of which was available in each participating institution. The arrangement of the sources within the phantom was prescribed as shown in Fig. 1 for the second trial, although not every institution decided to adopt this geometry in the first trial. With the more stringent protocol adopted for Trial 2, this arrangement was mandatory in order to eliminate the influence of apparent overlap of the activity of the sources.

The four sources were mounted in the phantom, which was then filled with water. The phantom was placed in the head position on the patient scanning couch and three sets of planar data and three sets of SPECT acquisitions (with CT, if available) were acquired and saved for processing. After making the measurements, participants were asked to report the details of their equipment, acquisition parameters, processing methods, and results.

For the second trial, a more prescriptive protocol was written, mostly dealing with harmonizing the acquisition parameters and the calibration methodology. Additional training was provided in the form of informal lectures given

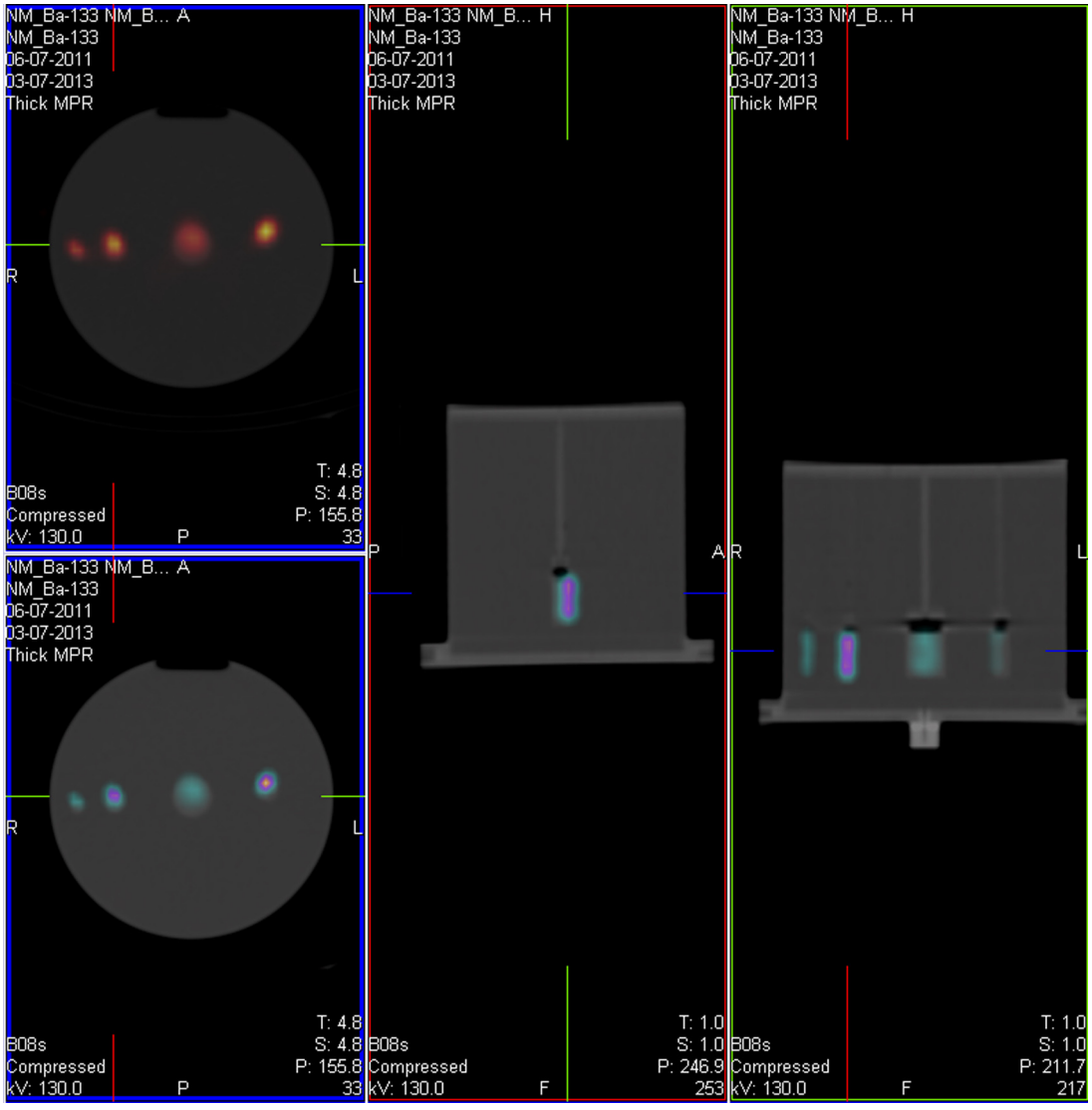

Fig. 1. SPECT-CT scan showing the prescribed placement of the 4 ^{133}Ba sources within the phantom.

by the more experienced participants in the project. The B-series source in air was again used for calibration, but the actual activity of only that source, as measured by NIST, was given to each participant – all other source activities remained unknown. For calibration, the acquisition and reconstruction parameters were the same as the data acquisitions for Trial 2 (see Table 2), but using the minimum practical radius. For the actual measurements, the participants were asked to perform three sets of SPECT acquisitions (with CT, if available). Further details on the acquisition and processing parameters are given in Table 2. The 5% lower scatter energy window was chosen so as to avoid interference with the 303 keV energy peak. In order to correct for spill-out the participants were asked either to draw the ROIs/VOIs such that they either included the majority of counts (planar, SPECT with Chang attenuation correction (Chang-AC) [21]) or they include the CT image of source dilated by 10 mm in all directions without overlap. The sites were also asked to repeat the respective measurement three times.

In all cases, the results to be reported were the details of their equipment, acquisition parameters, processing methods, the total counts of the acquisition in the peak window, and the estimated activity in each source for each acquisition.

Some sites lacked the ability to acquire transmission images needed for conjugate-view planar activity quantification method [22–25] and therefore estimated the object thickness using other methods. SPECT quantification without CT-based attenuation correction (CTAC) used the zero-order Chang correction method [21], where an estimate of the boundary of the phantom is required.

In order to exclude site-specific effects caused by operator-dependent data processing an additional centralized data analysis (CDA) was performed. For this purpose, the raw data from each site were collected by a central site (the participating

Table 2
Acquisition and reconstruction parameters for Trial 2.

General parameters			
Collimator	High energy		
Windowing	Peak	356 keV	15% Window
	Lower scatter	321 keV	5% Window
	Upper scatter	403 keV	10% Window
Matrix size	128 × 128		
Scan set-up	Planar	SPECT	SPECT-CT
Acquisition parameters			
Scan duration	600 s		60 s per view Step and shoot 120–128 total views
Reconstruction and post-processing parameters			Autocontour enabled, if available OSEM with at least 50 updates No post-reconstruction filter
Attenuation correction methods	Conjugate views or transmission measurements, if available. Attenuation coefficient to be used for ^{133}Ba: 0.111 cm^{-1}	Chang-AC	CTAC
Scatter correction parameters	1.66875 (lower window) 0.6675 (upper window)	Triple Energy Window Method, window weighting factors as given by the manufacturer	
ROI/VOI size	ROI/VOI to be drawn to include the majority of counts	CT image of source and dilated 1cm without overlap	

institute from Croatia) and processed following the prescriptive protocol established for the second trial (i.e. using the same reconstruction settings, corrections, and region drawing techniques for all data sets).

2.4 Analysis methodology

To facilitate consistent comparison of the results between the centres for each of the three exercises, all activity values reported by the participants are presented as ratios, R, of the mean of the participants' reported activities for each source at the prescribed reference time to the NIST calibrated activity for that source at the same reference time. Unless otherwise stated, the quoted uncertainties are standard ($k = 1$) uncertainties calculated from the quadratic addition of the relative combined standard uncertainty on the NIST activity for the specific source and the relative standard deviation of the three activity determinations for that source as reported by the participant.

Data were analysed for each of the three exercises separately by grouping the R-values by source size, thereby giving four subsets of data in each exercise. Within each subgroup, possible outliers were detected using Grubbs' test at a significance level of 5%. Each set was tested for normality after removal of outliers using a normal probability plot correlation coefficient test [26] at the 5% significance level. In each case, the data were indicated to be normally distributed.

For the calculation of a final central value for R within each exercise (see Table 3), mean values were first calculated, along with their standard deviations. It was apparent from the large variances that although suspected outliers (maximum of 2 in each set) were removed from the data sets, the means were still highly influenced by the presence of extreme values. For this reason, medians are also reported, since they tend to be more robust when one is confronted with an inconsistent data set. The standard uncertainties on the medians were calculated as median absolute deviations (MAD) [27].

One important aspect of the analysis of responses in the first exercise was that some of the participants reported the activity in terms of activity of ^{131}I instead of ^{133}Ba as instructed. In those cases, the participants' values were converted to equivalent ^{133}Ba activity by dividing by their measured relative response of their activity calibrator for the two radionuclides. For the analysis of data from Trial 1, an adjustment was made to account for differences between the participants' activity calibrator results and the NIST-calibrated activity. No such adjustment was necessary in Trial 2, as the true activity for the 4 mL source was provided to the participants.

Because it was not a goal of the exercise to evaluate individual institution performance, the results are presented in this paper in aggregate form, that is, the participant identifier does not necessarily correspond to the same institution in every plot.

3 Results

A summary of the average R values obtained for all the cylinders in each trial using each of the three methods is given in Table 3, along with their respective standard uncertainties. The results of each trial are presented in more detail below.

Table 3
Mean and median values of the ratio, R, of the participants' reported activities to the NIST-calibrated activity for all of the ^{133}Ba sources as measured by each technique for each trial. The uncertainties correspond to the standard deviations on the mean values or the median absolute deviations, in the case of the medians.

Method	Trial 1		Trial 2		CDA	
	Mean	Median	Mean	Median	Mean	Median
Planar						
Conjugate View	0.84(12)	0.86(7)	0.91(14)	0.88(8)	1.06(8)	1.06(6)
AC					1.09(4)	1.05(10)
SPECT						
Chang-AC	1.09(21)	1.10(10)	1.18(39)	0.98(10)	0.94(6)	0.94(4)
CTAC	1.08(13)	1.12(6)	1.06(8)	1.05(5)	1.00(8)	1.02(6)

Chang-AC is not widely used for quantitative SPECT; detailed results for this part of the trial are presented only in summary form. Hereafter, we focus solely on the results of planar imaging and of SPECT-CT.

3.1 Trial 1

The R values for the 4 cylindrical sources obtained from planar imaging, along with their associated standard uncertainties, are presented for each of the participants in Fig. 2a. A Grubb's test on the entire set of 32 values indicated that the two responses at $R = 3.19$ and $R = 1.54$ were indeed outliers and were not included in further calculations. The most striking result from this particular exercise is the fact that the majority of the measurements underestimate the activity by about 15%, independent of source size or institution. The responses ranged from $R_{min} = 0.53$ to $R_{max} = 1.04$.

In 5 of the data sets, an apparent trend towards greater underestimation in the ^{133}Ba activity is seen as the size of the source increases, which could be taken to be an attenuation effect. An analysis of variance (ANOVA) indicates, however, that there is no statistical difference ($p = 0.62$) between the results for the different source groups. The data do indicate that statistically significant differences in performance between the group of participants ($p = 0.031$) exist, but this is not unexpected, given the different levels of experience in each institution.

For the SPECT-CT measurements, shown in Fig. 2b, the data set was free of outliers, but had a relatively large range in responses ($R_{min} = 0.85$ to $R_{max} = 1.27$). No trends with regard to source size were detected and no statistically significant differences ($p = 0.18$) between laboratories were identified by ANOVA. Both the mean and median indicate an overestimation of the activity by about 10% for all the sources.

3.2 Trial 2

In addition to the data received from the laboratories participating in Trial 1, the data from Trial 2 included responses for both CTAC and planar from one participant that did not provide data in Trial 1. This gives a total of 9 institutes reporting results for planar and 5 institutes giving values for CTAC. It should also be noted that, as seen in Table 1, one of the participants used a scanner for this trial that was different from the one used in Trial 1. It is therefore possible that there are additional components of uncertainty when comparing the results for that participant that are not present in the others.

In the case of planar imaging, two results were identified by Grubb's tests to be outliers, with one value being too low ($R = 0.35$) and the other being too high ($R = 1.38$). It should be noted that in both cases, the results were for the smallest source, although there does not appear to be an influence of the source diameter on the measurement results across the entire data set.

Fig. 3a shows the planar imaging responses from each participant, including outliers, in Trial 2 plotted against their corresponding result in Trial 1. A clear difference in the performance between trials is not immediately evident from the data in Table 3, and, in fact, the ANOVA indicates only a marginal probability ($p = 0.04$) for there being a statistical difference between the two trials. Despite the fact that the means and medians did not shift significantly between the trials, it can be seen that the ranges shifted somewhat, from ($R_{min} = 0.53$, $R_{max} = 1.04$) in Trial 1 to ($R_{min} = 0.68$, $R_{max} = 1.19$), not including the outliers.

The participants' responses for Trial 2 using CTAC are plotted against the corresponding results in Trial 1 in Fig. 3b. What is immediately obvious in the plot is the fact that in both trials, this technique results in a general overestimation of the activity for all sources. No statistically significant difference between the trials is indicated ($p = 0.63$), other than it appears that the variability (standard deviation or MAD), was slightly lower in Trial 2.

3.3 Centralized data analysis (CDA)

The CDA was performed using a workflow provided by one of the authors (M. Lassmann) and implemented on a Siemens Syngo workstation.

The results of the CDA for the planar data are plotted against the participants' original responses for Trial 2 in Fig. 4a. It

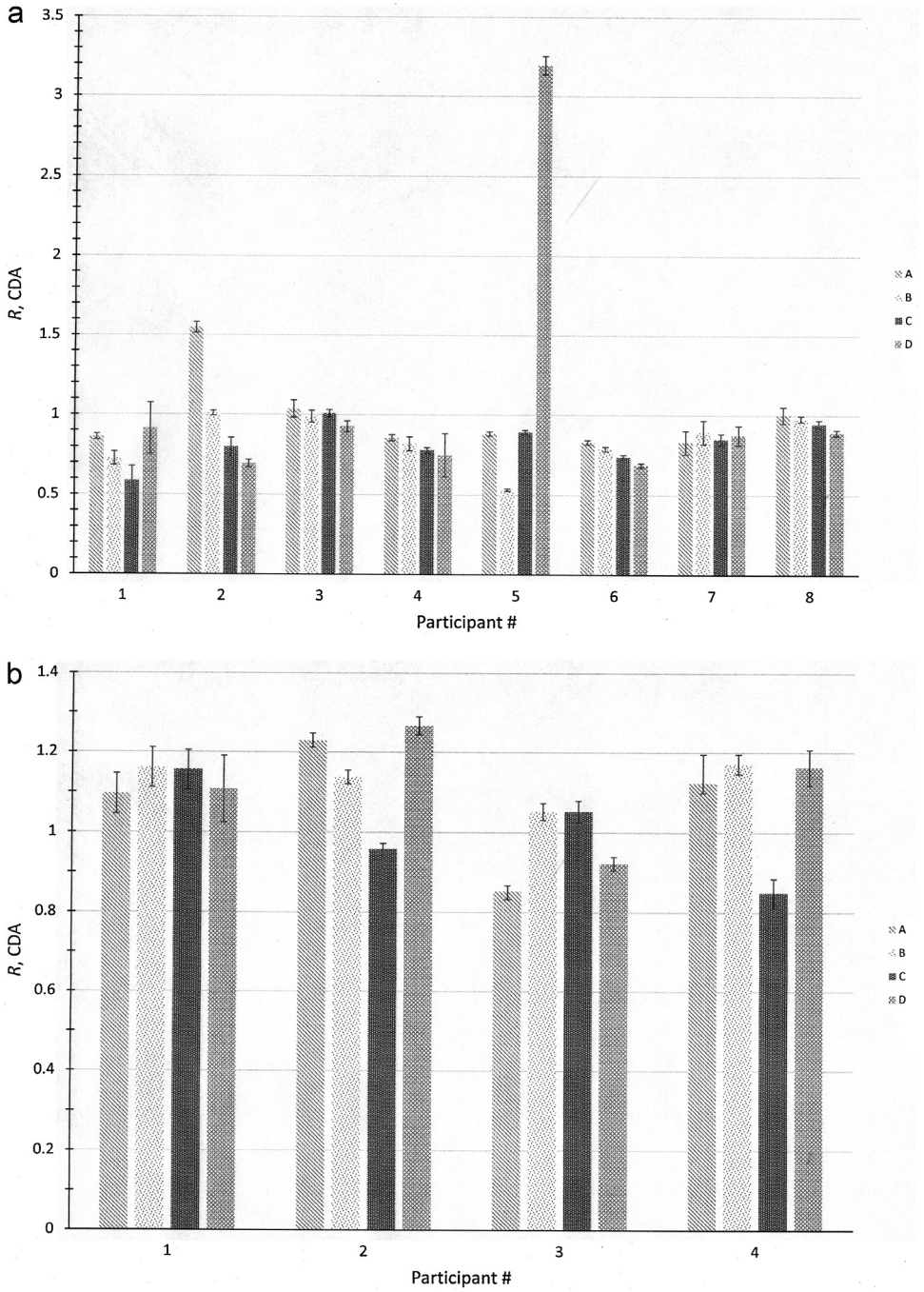

Fig. 2. (a) Plot of ratios, *R*, of participants' reported activities using planar imaging to the NIST-calibrated activity for the 4 sets (A through D) of ^{133}Ba sources in Trial 1. The uncertainty bars correspond to standard uncertainties and were calculated from the quadratic combination of the relative standard uncertainty on the NIST calibration for the respective source and the relative standard deviation on 3 repeated activity determinations of each source. (b) Plot of ratios, *R*, of participants' reported activities using SPECT-CT to the NIST-calibrated activity for the 4 sets (A through D) of ^{133}Ba sources in Trial 1. The uncertainty bars correspond to standard uncertainties and were calculated from the quadratic combination of the relative standard uncertainty on the NIST calibration for the respective source and the relative standard deviation on 3 repeated activity determinations of each source.

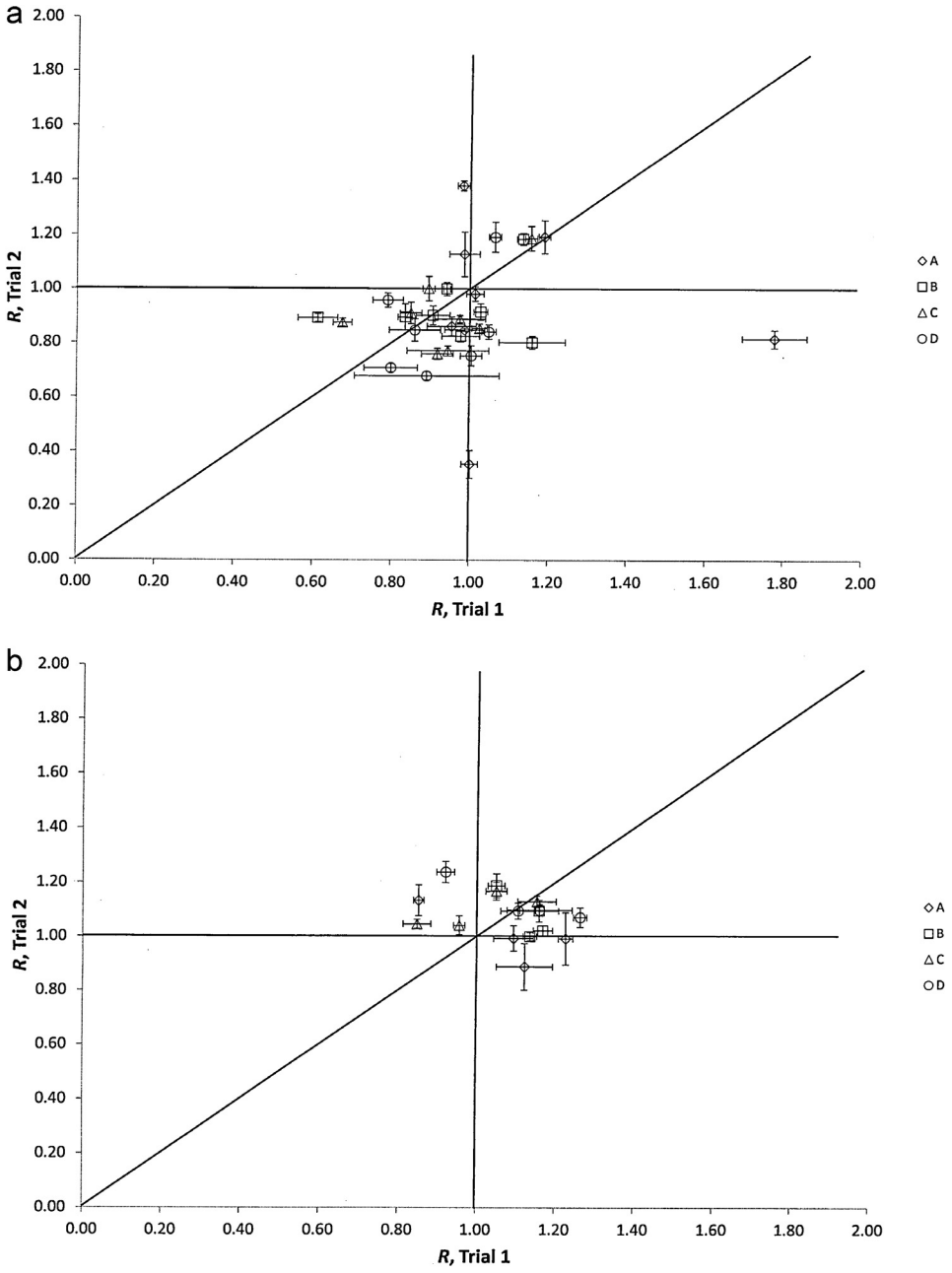

Fig. 3. (a) Plot of ratios, *R*, of participants' reported activities using planar imaging to the NIST-calibrated activity for the 4 sets (A through D) of ^{133}Ba sources for Trials 1 and 2. The uncertainty bars correspond to standard uncertainties and were calculated from the quadratic combination of the relative standard uncertainty on the NIST calibration for the respective source and the relative standard deviation on 3 repeated activity determinations of each source. The unity line is provided for guidance and would indicate those data for which the values in both exercises were equal. (b) Plot of ratios, *R*, of participants' reported activities using CTAC to the NIST-calibrated activity for the 4 sets (A through D) of ^{133}Ba sources for Trials 1 and 2. The uncertainty bars correspond to standard uncertainties and were calculated from the quadratic combination of the relative standard uncertainty on the NIST calibration for the respective source and the relative standard deviation on 3 repeated activity determinations of each source. The unity line is provided for guidance and would indicate those data for which the values in both exercises were equal.

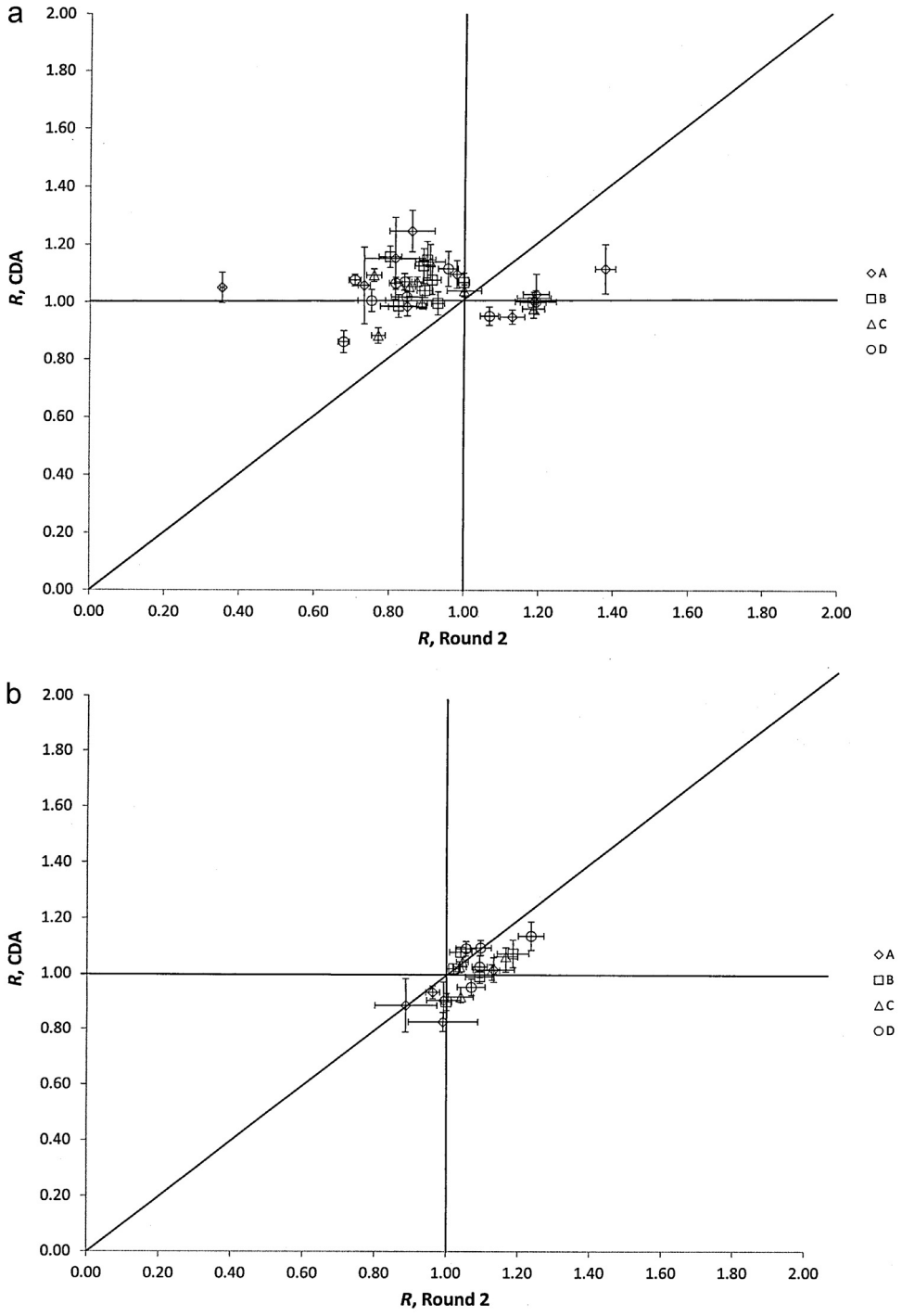

Fig. 4. (a) Plot of ratios, *R*, of participants' reported activities using planar imaging to the NIST-calibrated activity for the 4 sets (A through D) of ^{133}Ba sources for Trial 2 and the CDA. The uncertainty bars correspond to standard uncertainties and were calculated from the quadratic combination of the relative standard uncertainty on the NIST calibration for the respective source and the relative standard deviation on 3 repeated activity determinations of each source. The unity line is provided for guidance and would indicate those data for which the values in both exercises were equal. (b) Plot of ratios, *R*, of participants' reported activities using SPECT-CT imaging to the NIST-calibrated activity for the 4 sets (A through D) of ^{133}Ba sources for Trial 2 and the CDA. The uncertainty bars correspond to standard uncertainties and were calculated from the quadratic combination of the relative standard uncertainty on the NIST calibration for the respective source and the relative standard deviation on 3 repeated activity determinations of each source. The unity line is provided for guidance and would indicate those data for which the values in both exercises were equal.

is immediately apparent that an overall improvement in the quantitative accuracy was achieved when the same procedure is applied to all the data by a single analyst. This is also reflected in the improvement of the means and medians between Trial 2 and the CDA (i.e., closer to $R = 1$), as seen in Table 3. The CDA was also successful in correcting the outliers that were originally identified in the previous trials and reduced the overall range in R values to a minimum of 0.86 and maximum of 1.25. The ANOVA also indicate a difference between the two sets of data ($p = 2 \times 10^{-5}$).

Fig. 4b presents a plot of the results of the CDA for the participants' CTAC data plotted against the original responses for Trial 2. From the data, it is possible to identify a slight downward shift in the R values after the CDA. This is also seen in the slight lowering of the mean and median CTAC values in Table 3, which is indicated from the ANOVA as being statistically significant ($p = 0.017$).

The R values from the CDA, which are taken as the final responses from the participants for each imaging technique, are given in Fig. 5a for planar and Fig. 5b for SPECT-CT. Included in Fig. 5a are the results of a separate analysis that investigated the effect of transmission scan-based attenuation correction [28].

The ANOVA indicates a statistically significant ($p = 9.8 \times 10^{-4}$) difference between the R values when calculated with the two techniques when the entire group of all sources is considered. From each of the individual laboratories, however, only the data from participants 3 and 8 showed statistically significant differences. Because of the small number of degrees of freedom in the analyses ($n = 4$ for each participant and each method), however, caution should be used in drawing conclusions about the relative effectiveness of the two techniques.

For SPECT-CT, a trend of increasing R value with increasing size of the source can be seen that was not apparent in the original Trial 1 data. This trend is, in fact, statistically significant ($p = 0.024$) and is most likely attributable to uncorrected partial volume effects. Comparing the response between all the participants, it appears that the average R values for the 4 sources reported by Participant 3 are consistently lower than those of the other 4 laboratories. The magnitude of this difference is large enough for the ANOVA to detect the effect of the differences in response, since the entire data considered together indicates a between-participant difference ($p = 0.022$), but the probability is lowered to $p = 0.35$ when the data for Participant 3 are removed. While a definite cause for this effect is extremely difficult to identify without additional data, the fact that this particular institute had somewhat less experience with SPECT-CT than the others makes this a factor to consider.

4 Discussion

This study is the first international comparison with the purpose of comparing image quantification processes for molecular radiotherapy. The parameters of running an IAEA project strongly influenced the number and distribution of participants in this exercise. In particular, it is recognised that it would have been desirable to have more participants with a wider range of makes and models of scanners. However, the logistics of carrying out such a study for the first time showed this to be an acceptable, manageable size. Despite the limited number of institutions, this study produced a significant amount of useful data.

A set of very accurately calibrated and traceable sealed sources containing ^{133}Ba as a surrogate for ^{131}I were used for this study. The motivation for this was to avoid errors associated with the filling of the phantoms and with making the activity measurements in each laboratory. An earlier study by the IAEA on dose calibrator measurements showed that a variation in the measurement of activity, especially when using activity calibrators, is to be expected [13]. The use of a solid pre-calibrated source did present the disadvantage that it excluded measurements with background activity, which might have been more relevant for comparison with the clinical situation.

Although the main photon emission of ^{133}Ba has almost the same energy as ^{131}I (356 keV vs 364 keV), it does not have problems associated with the contribution from the 637 keV and 723 keV photons. Although the abundances are relatively low (7.3% and 1.8% respectively), these photons are highly penetrating and contribute to the events in the 364 keV window with a fraction up to 20% [18]. In the case of imaging with radionuclides where septum penetration is significant, such as ^{131}I, the definition of the sensitivity is even more important. Determining whether or not the high-energy photons have a strong influence on the quantification when comparing results between studies with ^{133}Ba and ^{131}I will require a more systematic investigation that must take into account all of the components of uncertainty in the imaging process.

The activities in the four sources were such that the dead-time for the imaging system was not an issue. For patient imaging, and especially when imaging after therapeutic doses, high count rates can be expected due to high administered activities. The effects in the imaging from count rate and pileup effect are camera specific, but generally result in a decreased measured count rate [29,30]. Therefore, activity is generally underestimated if a sensitivity factor measured at a low count rate is used. Very high count-rates can also introduce artefacts in the image due to the way events are positioned (i.e. from the centroid of the scintillation light) [31].

The sensitivity factor relates the count rate to the activity. Experience from this project showed that the sensitivity calculated from the CDA was site-dependent. The deviations were, however, small with a tendency to higher sensitivities for the SPECT-CT studies. It is of importance for future studies that the method for the sensitivity calculation be well defined and documented [32], as the sensitivities reported were

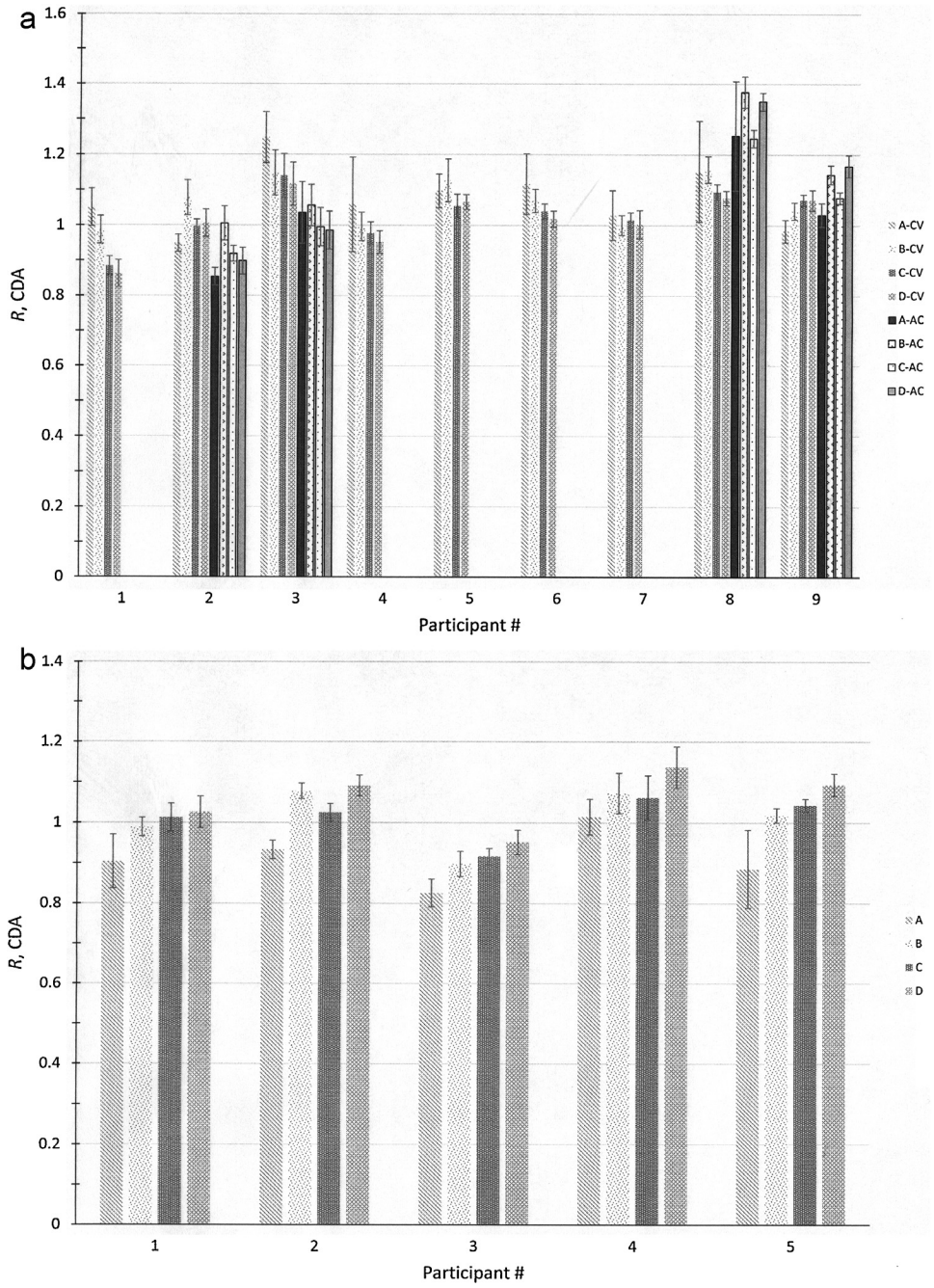

Fig. 5. (a) Plot of ratios, *R*, of participants' reported activities using planar imaging with Conjugate View (CV) and Attenuation Correction (AC) to the NIST-calibrated activity for the 4 sets (A through D) of ^{133}Ba sources from the CDA. The uncertainty bars correspond to standard uncertainties and were calculated from the quadratic combination of the relative standard uncertainty on the NIST calibration for the respective source and the relative standard deviation on 3 repeated activity determinations of each source. (b) Plot of ratios, *R*, of activities from the CDA with CTAC to the NIST-calibrated activity for the 4 sets (A through D) of ^{133}Ba sources. The uncertainty bars correspond to standard uncertainties and were calculated from the quadratic combination of the relative standard uncertainty on the NIST calibration for the respective source and the relative standard deviation on 3 repeated activity determinations of each source.

sometimes inconsistent. For one site using a camera with a crystal thickness of 15.9 mm, the sensitivity was about 60% higher as compared to the sensitivity of cameras of the same vendor using the standard crystal thickness of 9.5 mm which, as a consequence, results in improved count statistics for this particular camera.

If a whole field-of-view sensitivity is used to convert counts per second to MBq in a narrow ROI, as was done for this comparison, then the activity might be underestimated due to spill-out of the counts caused by the limited spatial resolution of the scintillation cameras. Therefore, we decided in Trial 2 that the ROIs/VOIs should not be drawn close to the boundaries of the object (e.g. as defined by CT) but larger in order to include spill-out (see Table 2). More sophisticated methods to correct for partial volume effects have been developed but are generally not clinically available [33].

In this project, with a phantom having a fairly simple source geometry, quantification using planar imaging worked reasonably well. No major changes in terms of overall accuracy were observed between Trial 1 and Trial 2, although the protocol settings were much more uniform for Trial 2. Obviously, the planar image quantification methods are still operator-dependent even if there is a uniform protocol for all sites. The lowest deviation was observed after the CDA of Trial 2. This is most likely caused by the fact (i) that the sensitivity factors were recalculated using a standardised procedure, (ii) the data were evaluated by a single observer applying the same attenuation and scatter method and ROI drawing to all data sets.

For SPECT studies, the Chang-AC method also performed reasonably well, with R values and uncertainties similar to those obtained with planar imaging. The setup for the correction involves several steps for the user that may not be straightforward. The outline of the object being quantified is normally manually defined by fitting a circular boundary. Due to the lack of background activity in the configuration used for these studies, the phantom boundary was not easily seen, and thus the outline was difficult to define. Some sites added ^{133}Ba external point sources to the surface of the phantom in order to be able to define the boundary with a protocol provided by the vendor.

The smallest deviation from the actual activity of the source was with SPECT-CT. Only slight improvement was achieved by the CDA of the data of Trial 2. In our study, we could consistently quantify the activity within 10% of the NIST-calibrated activity with an appropriate quantification procedure. There might be, however, a vendor-specific bias of this study as most SPECT-CT scanners used for this trial were from a single manufacturer. In addition, we observed for the SPECT-CT studies only a consistent source-volume dependent increase of the activity ratios that might be caused by partial volume effects, particularly for the 2 mL and 4 mL sources. It should be stressed that although the simplified geometry and acquisition conditions enabled the errors on the quantification for the smallest sources to be typically around 10% (with a best case of 2%), this level of accuracy is probably not easily achievable clinically with the current state of practice.

As specified in the protocol, the participants were asked to repeat the same study three times, thereby allowing inherent site-specific variability to be assessed. The standard deviations within one study and one site were, for each trial, much smaller than the variation of results between the sites. This indicates that modern scintillation cameras provide stable data for repeated studies [34,35].

A few sites were able to perform attenuation correction using transmission measurements. Although there were not enough data to be able to draw meaningful conclusions about the quantitative accuracy of this technique, it did become apparent during the data analysis phase that the choice of attenuation coefficient is critical for accurate and consistent results.

After evaluating the results of the study it is our assertion that, for quantitative imaging, if SPECT-CT is available, this modality is to be recommended because of the smaller number of manual procedures (and thus less chance for errors induced by the operator) and also the fact that it supports both uniform and non-uniform attenuation corrections.

It should be pointed out that the phantom and source distribution used in this study does not allow for a fair comparison between quantitative imaging using SPECT-CT and planar imaging. There are well-known problems associated with planar imaging including the unwanted contribution of events from over- and underlying activities, self-attenuation in distributed source volumes and the problems of measuring attenuation factors for planar quantitation. These factors are likely to decrease the accuracy and increase the variability of planar quantitation while having a smaller effect on SPECT-CT quantitation. Thus, SPECT-CT-based quantitation is likely to demonstrate even greater superiority in such cases. However, this paper does not provide direct evidence of this, and such a comparison was beyond the scope of this work.

The project demonstrated the need for training and standardized protocols to achieve good levels of quantitative accuracy and repeatability in a multicentre setting. In general, the need for standardization of education and training of medical physicists is well recognized by the IAEA; a number of publications have been produced to fill this gap [16,36–38].

Since the time that this project was started, other national and regional efforts, such as the recently completed Metrology for Molecular Radiotherapy (MetroMRT) project [39] have been initiated with many of the same goals, but with slightly different methods for evaluating image quantification performance. It is hoped that with the publication of such data, improvements in the state of quantitative imaging can be proposed and moved into clinical practice.

5 Conclusions

This work underlines several facts concerning quantitative imaging for molecular radiotherapy:

- Solid barium-133 sources can be used for testing image quantification procedures in multicentre trials thus avoiding inherent problems with on-site activity measurement and phantom preparation.
- A detailed protocol for calibration measurements is needed.
- Large uncertainties were found for planar methods and the studies that applied Chang's attenuation correction. This was most likely caused by the difficulties encountered in attenuation correction and ROI drawing.
- SPECT-CT showed the best accuracy and lowest variability due to the built-in inherent and theoretically rigorous correction methods. In addition, SPECT-CT requires less interaction by the operator when compared to other methods.
- There were much smaller variations at the same site than across the different sites.
- For multicentre studies there is a need for a consistent and specific description of each step in the acquisition and processing chain in order to achieve accurate and reliable image quantification.
- Although the simplistic phantom design and lack of background provided excellent data on the potential accuracy limits for SPECT and planar image quantification, such results are probably not yet achievable clinically with current methodologies.

Overall this study showed, for the first time, that international multicentre trials for image quantification in molecular radiotherapy can be carried out and may be the first step towards improved traceability of dosimetry studies in patients.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation by the National Institute of Standards and Technology or the International Atomic Energy Agency, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Acknowledgements

The authors would like to thank RadQual, LLC for providing the sources used in this study.

Dr. Stig Palm (formerly IAEA, presently Gothenburg University) is acknowledged for conceiving and initiating this project.

This work was carried out as part of the IAEA Coordinated Research Project E2.10.07 entitled "Development of Quantitative Nuclear Medicine Imaging for Patient Specific Dosimetry".

References

[1] Loevinger R, Budinger TF, Watson EE. MIRD Primer for Absorbed Dose Calculations (Revised Edition). The Society of Nuclear Medicine, MIRD; 1991.

[2] Iida H, Nakagawara J, Hayashida K, Fukushima K, Watabe H, Koshino K, et al. Multicenter evaluation of a standardized protocol for rest and acetazolamide cerebral blood flow assessment using a quantitative SPECT reconstruction program and split-dose ^{123}I-iodoamphetamine. J Nucl Med 2010;51(10):1624–31.

[3] Macey DJ, DeNardo GL, DeNardo SJ. Planar gamma camera quantitation of ^{123}I, ^{99m}Tc or ^{111}In in the liver and spleen of an abdominal phantom. Cancer Biother Radiopharm 1999;14(4):299–306.

[4] Yoneda H, Shirao S, Koizumi H, Oka F, Ishihara H, Ichiro K, et al. Reproducibility of cerebral blood flow assessment using a quantitative SPECT reconstruction program and split-dose ^{123}I-iodoamphetamine in institutions with different gamma-cameras and collimators. J Cereb Blood Flow Metab 2012;32(9):1757–64.

[5] Slomka PJ, Berman DS, Germano G. Absolute myocardial blood flow quantification with SPECT/CT: is it possible? J Nucl Cardiol 2014;21(6):1092–5.

[6] Glatting G, Lassmann M. Nuclear medicine dosimetry: quantitative imaging and dose calculations. Z Med Phys 2011;21(4):246–7.

[7] Hanscheid H, Lassmann M, Luster M, Thomas SR, Pacini F, Ceccarelli C, et al. Iodine biokinetics and dosimetry in radioiodine therapy of thyroid cancer: procedures and results of a prospective international controlled study of ablation after rhTSH or hormone withdrawal. J Nucl Med 2006;47(4):648–54.

[8] Sanders JC, Kuwert T, Hornegger J, Ritt P. Quantitative SPECT/CT imaging of ^{177}Lu with in vivo validation in patients undergoing peptide receptor radionuclide therapy. Mol Imaging Biol 2015;17(4):585–93.

[9] Bardies M, Buvat I. Dosimetry in nuclear medicine therapy: what are the specifics in image quantification for dosimetry? Q J Nucl Med Mol Imaging 2011;55(1):5–20.

[10] Council Directive 2013/59/Euratom of 5 Dec. 2013, Laying Down Basic Safety Standards for Protection against the Dangers Arising from Exposure to Ionising Radiation, and Repealing Directives 89/618/Euratom, 90/641/Euratom, 96/29/Euratom, 97/43/Euratom and 2003/122/Euratom, Basic Safety Standard, L13; 2014.

[11] Bailey DL, Willowson KP. Quantitative SPECT/CT: SPECT joins PET as a quantitative imaging modality. Eur J Nucl Med Mol Imag 2014;41(Suppl. 1):S17–25.

[12] Berker Y, Goedicke A, Kemerink GJ, Aach T, Schweizer B. Activity quantification combining conjugate-view planar scintigraphies and SPECT/CT data for patient-specific 3-D dosimetry in radionuclide therapy. Eur J Nucl Med Mol Imag 2011;38(12):2173–85.

[13] International Atomic Energy Agency. Quality Assurance for Radioactive Measurement in Nuclear Medicine. IAEA Technical Reports Series No 454; 2006.

[14] Dewaraja YK, Frey EC, Sgouros G, Brill AB, Roberson P, Zanzonico PB, et al. MIRD pamphlet No. 23: quantitative SPECT for patient-specific 3-dimensional dosimetry in internal radionuclide therapy. J Nucl Med 2012;53(8):1310–25.

[15] Dewaraja YK, Ljungberg M, Green AJ, Zanzonico PB, Frey EC, Bolch WE, et al. MIRD pamphlet No. 24: guidelines for quantitative ^{131}I SPECT in dosimetry applications. J Nucl Med 2013;54(12):2182–8.

[16] International Atomic Energy Agency. Quantitative Nuclear Medicine Imaging: Concepts, Requirements and Methods. IAEA Human Health Reports No 9; 2014.

[17] Bé M, Chisté V, Dulieu C, Browne E, Chechev V, Kuzmenko N, et al. Table of radionuclides (Vol. 1-A = 1 to 150). Bureau International des Poids et Mesures; 2004.

[18] Dewaraja YK, Ljungberg M, Koral KF. Characterization of scatter and penetration using Monte Carlo simulation in ^{131}I imaging. J Nucl Med 2000;41(1):123.
[19] Zimmerman BE, Pibida L, King LE, Bergeron DE, Cessna JT, Mille MM. Calibration of traceable solid mock I-131 phantoms used in an international SPECT image quantification comparison. J Res Natl Inst Stand Technol 2013;118:359–74.
[20] Ogawa K, Harata Y, Ichihara T, Kubo A, Hashimoto S. A practical method for position-dependent Compton-scatter correction in single photon emission CT. IEEE Trans Med Imaging 1991;10(3):408–12.
[21] Chang L-T. A method for attenuation correction in radionuclide computed tomography. IEEE Trans Nucl Sci 1978;25(1):638–43.
[22] Shulkin BL, Sisson JC, Koral KF, Shapiro B, Wang XH, Johnson J. Conjugate view gamma camera method for estimating tumor uptake of iodine-131 metaiodobenzylguanidine. J Nucl Med 1988;29(4):542–8.
[23] Thomas SR, Maxon HR, Kereiakes JG. In vivo quantitation of lesion radioactivity using external counting methods. Med Phys 1976;03(04):253–5.
[24] Fleming JS. A technique for the absolute measurement of activity using a gamma camera and computer. Phys Med Biol 1979;24(1):176–80.
[25] Macey DJ, Marshall R. Absolute quantitation of radiotracer uptake in the lungs using a gamma camera. J Nucl Med 1982;23(8):731–4.
[26] Filliben JJ. The probability plot correlation coefficient test for normality. Technometrics 1975;17(1):111–7.
[27] Muller JW. Possible advantages of a robust evaluation of comparisons. J Res Natl Inst Stand Technol 2000;105(4):551–6.
[28] Siegel JA, Thomas SR, Stubbs JB, Stabin MG, Hays MT, Koral KF, et al. MIRD pamphlet no. 16: techniques for quantitative radiopharmaceutical biodistribution data acquisition and analysis for use in human radiation dose estimates. J Nucl Med 1999;40(2):37S–61S.
[29] Chiesa C, Negri A, Albertini C, Azzeroni R, Setti E, Mainardi L, et al. A practical dead time correction method in planar activity quantification for dosimetry during radionuclide therapy. Q J Nucl Med Mol Imaging 2009;53(6):658–70.
[30] Hobbs RF, Baechler S, Senthamizhchelvan S, Prideaux AR, Esaias CE, Reinhardt M, et al. A gamma camera count rate saturation correction method for whole-body planar imaging. Phys Med Biol 2010;55(3):817–31.
[31] IAEA. Quality Control Atlas for Scintillation Camera Systems. International Atomic Energy Agency; 2003.
[32] Lassmann M, Chiesa C, Flux G, Bardiès M. EANM dosimetry committee guidance document: good practice of clinical dosimetry reporting. Eur J Nucl Med Mol Imaging 2011;38(1):192–200.
[33] Erlandsson K, Buvat I, Pretorius PH, Thomas BA, Hutton BF. A review of partial volume correction techniques for emission tomography and their applications in neurology, cardiology and oncology. Phys Med Biol 2012;57(21):R119–59.
[34] Anizan N, Wang H, Zhou XC, Hobbs RF, Wahl RL, Frey EC. Factors affecting the stability and repeatability of gamma camera calibration for quantitative imaging applications based on a retrospective review of clinical data. Eur J Nucl Med Mol Imag Res 2014;4(1):67.
[35] Anizan N, Wang H, Zhou XC, Wahl RL, Frey EC. Factors affecting the repeatability of gamma camera calibration for quantitative imaging applications using a sealed source. Phys Med Biol 2015;60(3):1325–37.
[36] IAEA. Nuclear Medicine Physics: A Handbook for Teachers and Students. International Atomic Energy Agency; 2015.
[37] International Atomic Energy Agency. Clinical Training of Medical Physicists Specializing in Nuclear Medicine. IAEA Training Course Series No 50; 2011.
[38] International Atomic Energy Agency. Postgraduate Medical Physics Academic Programmes. IAEA Training Course Series No 56; 2013.
[39] D'Arienzo M, Capogni M, Smyth V, Cox M, Johansson L, Solc J, et al. Metrological issues in molecular radiotherapy. In: EPJ Web of Conferences: EDP Sciences. 2014. p. 00022.