University of Groningen

Standardisation and harmonisation of quantitative oncology PET/CT studies

Kaalep, Andres

DOI:
[10.33612/diss.122708993](https://doi.org/10.33612/diss.122708993)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Kaalep, A. (2020). *Standardisation and harmonisation of quantitative oncology PET/CT studies*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen.
<https://doi.org/10.33612/diss.122708993>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Standardisation and harmonisation of quantitative oncology PET/CT studies

Andres Kaalep

Colophon

Printed by Gildeprint, Enschedé

Designed by Nina Kleingeld | Stikstof Studio

ISBN 978-94-6402-171-4

Copyright © 2020 Andres Kaalep

All rights reserved - no part of this thesis may be reproduced or transmitted in any form or by any means without prior written permission of the copyright owner.

university of
 groningen

Standardisation and harmonisation of quantitative oncology PET/CT studies

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 20 April 2020 at 12.45 hours

by

Andres Kaalep

born on 31 July 1985
in Eesti, Estland

Supervisor

Prof. R. Boellaard

Co-supervisors

Prof. J.R. de Jong

Prof. T. Sera

Assessment Committee

Prof. J. Pruim

Prof. S. Stroobants

Prof. H.W.A.M. de Jong

Table of Contents

1. General Introduction	9
1.1 History	11
1.2 Positron emission tomography (PET)	13
1.2.1 Corrections to raw data	15
1.2.2 Image reconstruction	16
1.2.3 Quantitative imaging in PET	18
1.3 Multicentre standardisation programs	21
1.4 Thesis aim	23
1.5 Thesis outline	23
Bibliography	24
2. EANM/EARL FDG-PET/CT accreditation - summary results from the first 200 accredited imaging systems	27
2.1 Abstract	29
2.2 Introduction	30
2.3 Materials and methods	31
2.3.1 Acquisition and submission of data to EARL	31
2.3.2 Quantitative analysis and approval by EARL	32
2.3.3 Data clean-up and preparation	33
2.3.4 Analysis	33
2.4 Results	34
2.4.1 General overview	34
2.4.2 Calibration QC	34
2.4.3 Image Quality QC	38
2.5 Discussion	38
2.6 Conclusion	41
2.7 Compliance with Ethical Standards	42
2.8 Supplementary material	43
Bibliography	48
3. Feasibility of state of the art PET/CT systems performance harmonisation	51
3.1 Abstract	53
3.2 Introduction	54
3.3 Materials and methods	55
3.3.1 PET/CT system selection	55
3.3.2 Phantom experiments	55

3.3.3	Acquisition & reconstruction parameters	56
3.3.4	Data analysis	58
3.3.5	Selection of harmonising reconstruction modes	58
3.3.6	Mean Contrast Recovery (MCR)	61
3.3.7	Contrast Recovery Variability (CRV)	61
3.3.8	Noise	61
3.3.9	Curvature and absolute error	63
3.3.10	Visual analysis	63
3.3.11	Validation of reconstruction modes for harmonisation	63
3.4	Results	63
3.4.1	New specifications proposed for harmonisation	63
3.4.2	Mean Contrast Recovery (MCR)	64
3.4.3	Contrast Recovery Variability (CRV)	64
3.4.4	Noise	65
3.4.5	Curvature and absolute error	65
3.4.6	Visual analysis	65
3.4.7	Additional reconstructions	76
3.5	Discussion	76
3.5.1	Limitations and future directions	78
3.6	Conclusions	80
3.7	Acknowledgments	80
3.8	Disclaimer	81
3.9	Compliance with Ethical Standards	81
3.10	Supplemental Material	82
	Bibliography	91

4. Quantitative implications of the updated EARL 2019

PET-CT performance standards	95
4.1 Abstract	97
4.2 Introduction	98
4.3 Materials and methods	100
4.3.1 Selecting post-filtering parameters by phantom experiments	100
4.3.2 Patient selection and preparation	100
4.3.3 Lesion selection, segmentation and analysis	102
4.4 Results	104
4.4.1 Phantom data	104
4.4.2 Clinical data	105
4.5 Discussion	114
4.6 Conclusions	116
4.7 Declarations	116

4.7.1	Abbreviations	116
4.7.2	Ethical Approval and Consent to participate	117
4.7.3	Consent for publication	117
4.7.4	Supplemental material	118
4.7.5	Competing interests	123
4.7.6	Funding	123
4.7.7	Authors' contributions	123
4.7.8	Acknowledgements	123
	Bibliography	124

5. Feasibility of PET/CT system performance harmonisation for quantitative multicenter ⁸⁹Zr studies 127

5.1	Abstract	129
5.2	Introduction	130
5.3	Materials and methods	130
5.3.1	Investigated systems & phantom experiments	130
5.3.2	Data analysis	132
5.4	Results	132
5.5	Discussion	137
5.6	Conclusions	138
5.7	Declarations	138
5.7.1	Ethics approval and consent to participate	138
5.7.2	Consent for publication	139
5.7.3	Availability of data and materials	139
5.7.4	Competing Interest	139
5.7.5	Funding	139
5.7.6	Authors contribution	139
5.7.7	Acknowledgments	140
	Bibliography	141

6. Future perspectives and concluding remarks 143

6.1	Future directions	145
6.2	Concluding remarks	147
	Bibliography	148

7. Summary of findings 151

	Nederlandse samenvatting	156
	Bibliography	159
	Curriculum Vitae	161
	Acknowledgements	163

Chapter 1

General Introduction

1.1 History

The origins of nuclear medicine can be traced back to the end of the 19th century when Henri Becquerel, while working on phosphorescence, discovered that some uranium salts spontaneously emit an invisible penetrating radiation, much like the X-rays discovered by Wilhelm Conrad Röntgen just a year earlier. A period of intense research into radioactivity followed which led to the discovery of additional radioactive elements of thorium, polonium and radium. In 1903 Marie Curie, Pierre Curie and Henri Becquerel jointly received the Nobel Prize in physics for their work in the field of radioactivity. Their discoveries were the basis for the field of nuclear medicine [1][2].

While the electron was discovered already in 1897 by Sir Joseph John Thomson studying the properties of cathode rays [3], it took another 35 years until its antiparticle, the positron, was first described by Carl Anderson as a distinct particle and evidence of antimatter. Investigating cosmic ray interactions in a cloud chamber, Anderson found in 1932 that in a magnetic field some particle trails were bent similarly to those of the electrons, but in the opposite direction. This could only mean a particle with the same mass but opposite charge – an antielectron or positron.

Although humanity has always been exposed to ionizing radiation, we have not developed any senses for its detection. With the discovery of X-rays and radioactivity, the need arose for methods and instruments to detect and study the new phenomenon. Initially photographic plates and emulsions were developed that made use of the darkening of the material under irradiation. In 1899, thanks to the work of Thomson, the operating characteristics of ionization chambers were well understood and consequently used by Marie Curie in her investigation into radioactive materials. The first instrument that could detect individual rays was the spinthariscopes, invented by William Crookes in 1903. The first cloud chamber was built in 1911 by Charles Thomson Rees Wilson which in addition to detecting individual rays, makes it possible to visualise the path of the ray or particle. In 1928 Hans Geiger and Walther Muller introduced the so called Geiger-Muller or GM counter which responded to individual radiation-induced events by giving a high level output signal. The GM counter found wide use due to its low cost, simplicity and ease of operation. The drawbacks were that it could not directly measure the energy of radiation and was limited to relatively low counting rates [4]. In 1948, Harmut Kallmann suggested that the scintillations produced when radiation interacted with certain types of materials could be individually detected and amplified by photomultiplier tubes for electronic counting [5]. This, along with the discovery of calcium tungstate by Benedict Cassen, Lawrence Curtis and Clifton Reed as a detector for high energy gamma photons, led to the development of the scintillator detector [6].

In 1951 Frank Wrenn, Myron Good and Philip Handler proposed a possible use of thallium-activated sodium iodide (NaI) detectors opposite to each other [7]. From experiments comparing single photon and coincidence detection, Wrenn concluded „thus, it appears possible to more accurately delimit point sources, and hence extended sources with the technique of coincidence counting of annihilation pairs”. Independently working Gordon Brownell and William Sweet reported in 1953 on a device designed for localization of brain tumours [8]. It had two collinear NaI detectors mounted on an adjustable platform moving in a rectilinear fashion and a printer recording the coincidence counting rate. The positron scanners developed at the time, although using coincidence detection instead of lead collimators, still lacked the required sensitivity.

In 1961, to improve sensitivity, Rankowitz et al. positioned a complete ring of scintillation detectors around the object to be scanned [9]. Still limited by under-sampling and lack of attenuation correction, the system demonstrated the advantages of electronic collimation. In 1963, Hal Anger and Alexander Gottschalk designed a prototype system capable of imaging an entire brain with no mechanical scanning motion or collimators. The authors reported a sensitivity about 20 times higher than what was achieved with a collimator system under identical conditions [10].

At the beginning of 1970s, Godfrey Hounsfield and James Ambrose proposed a method of using radiograph transmission measurements at multiple angles through the head of a patient. A computer would calculate the absorption values and display these as tomographic slices [11–13].

In the years 1972 – 1973 Michael Phelps, Edward Hoffman, Nizar Mullani and Michel Ter-Pogossian from Washington University built a device called PETT II, which used annihilation coincidence detection to reconstruct transaxial tomograms [14]. Reconstruction was done using the locally developed Fourier-based method instead of the algebraic base method used on CT scanners. Attenuation correction measurements were done using a thin plastic ring filled with ^{64}Cu placed around the object. In 1975, the prototype was developed into a clinically applicable PET (III) whole-body camera for which the sensitivity allowed to perform scanning in 2–4 minutes per slice [15]. The PET (III) system was used in Washington University and later at Brookhaven National Laboratory for both human and animal studies.

In the late 1970s and 1980s the single ring of NaI scintillation crystals was expanded to multiple rings and NaI replaced with a more suitable bismuth germanate oxide (BGO) [16]. The use of BGO represented a major advancement over NaI because the twice higher density allowed the use of smaller crystals while having three times the detection efficiency. Another type of scintillator crystal, cesium fluoride was also investigated, due to its fast scintillation and application in time-of-flight detection schemes [17], leading to the eventual development of commercial time-of-flight detectors.

In 1990s, most systems used BGO crystals and larger axial field of view (>15 cm) with 4-5 mm spatial resolution. Also 3D imaging became available by removing the lead septa separating the detector rings so all detectors in all planes can register coincidence events [18].

All of the previously mentioned inventions and discoveries have contributed significantly to the imaging systems we see in use today. Most recent generation of positron emission tomography systems have no photomultiplier tubes and use digital detectors coupled directly to scintillator crystals instead, resulting in increased spatial and timing resolution as well as highly stable performance across a large dynamic range of count rates [19]. While BGO is still used as a scintillator material, new lutetium based fast scintillators are widely utilised to facilitate better time-of-flight characteristics of modern PET systems. Axial field of view has increased to >25 cm and with it the sensitivity to >22 cps/kBq. A total body scanner is already in development. All contemporary PET systems by major vendors are equipped with a combined CT or MR modality used for attenuation correction and additional anatomical and functional information.

1.2 Positron emission tomography (PET)

PET is a functional imaging technique in nuclear medicine where radiopharmaceutical is administered intravenously to a patient and the behaviour of the substance within the body is monitored using specialised imaging equipment. A radiopharmaceutical consists of a biologically active molecule, called tracer, and a radionuclide that in case of PET, is a positron emitting radioisotope which allows the visualisation of the radiolabelled tracer in a PET system.

Many different tracer molecules have been developed which are used to observe various metabolic processes such as glucose metabolism, tumour growth, blood flow, receptor expression [20][21][22]. After administration to the patient, the radiolabelled compound travels throughout the bloodstream and accumulates based on the biological properties of the tracer. In the case of the glucose analogue 2-deoxy-2- (^{18}F) fluoro-D-glucose (^{18}F -FDG) glucose utilising parts of the body. In the patient's body the nucleus of the radioisotope used as a label decays by emitting a positron. The emitted positron travels a small (≤ 3 mm, in case of $[^{18}\text{F}]$) distance losing its remaining kinetic energy, after which it interacts with an electron, both annihilate resulting in two 511 keV photons emitted essentially in opposite directions. A PET camera looks for these simultaneously emitted photons and when two are detected within the specified timing and energy windows, a coincidence event is registered and the annihilation is assumed to have taken place along a line, called line-of-response (LOR), connecting the two detectors that registered the photons. When time-of-flight acquisition technology is used,

Figure 1. Types of coincidence detection in a PET system.

the time interval between the two arriving coincidence photons is measured, which allows the positioning of the annihilation event on the LOR to within a couple of centimetres. Due to the limited transaxial field of view (FOV), the acquisition is limited to a small (usually ≤ 25 cm) region at a time and the whole examined area is covered by either a series of separate or one continuously advancing acquisition across the area of interest. During the acquisition millions of coincidence events called prompts are collected which can originate from either a *true*, *scattered* or a *random* event (Fig. 1). In a *true coincidence* event, both of the annihilation photons travel in a straight line to the PET detectors where they are registered. A *scatter event* is caused by one or both annihilation photons changing their trajectory and losing energy due to Compton scattering, while a *random event* is caused by two unrelated annihilation photons arriving to the detectors within the specified timing window erroneously causing a coincidence event to be registered. Both scatter and random events cause false LORs to be detected and need to be corrected for by algorithms calculating scatter contribution and estimating the number of randoms by singles count rate or a delayed timing window. Acquired PET data needs to be corrected for attenuation of the photons within the patient's body. For this, modern PET scanners accommodate a CT part, making them hybrid PET/CT scanners. The CT data acquired just before or after PET scan and in the same patient position, can be used to assign attenuation coefficients to all voxels within the investigated volume to create an attenuation map and correct for loss of signal due to attenuation within the patient. More recently, advances in PET detectors have allowed hybrid PET/MR systems to be created, which provide good soft tissue contrast from the MR and lower overall radiation dose by skipping the CT altogether. PET/MR has however some unresolved issues due to lacking good attenuation correction within certain regions of the body. Additionally, the more complicated installation requirements and increased cost make PET/MR systems less popular than conventional PET/CTs.

1.2.1 Corrections to raw data

The abovementioned physical and technical issues unavoidably affect the acquired data and cause bias and artefacts in the acquired images. In order that the final image data be suitable for accurate visual and quantitative evaluation, corrections need to be applied to the raw data before and during image reconstruction.

Attenuation correction was historically performed by irradiating the scanner FOV with a circular or a rotating radioactive source. These transmission measurements were used to calculate the attenuation coefficients within the FOV. In some cases atlas based attenuation coefficients may be used. All modern PET/CT systems utilise the integrated CT imaging capabilities that allow for a large photon flux in a short acquisition time to provide PET with an accurate attenuation map. For the emerging MR/PET systems, the attenuation coefficients need to be derived from an MR images, which requires conversion of MR data to 511keV attenuation coefficients. The accuracy of MR based attenuation correction remains one of the main issues limiting the use in areas of the body where synthetic attenuation maps are more difficult to acquire due to no or minimal signal (i.e. lungs) [23–26].

Random coincidences are inevitably collected during PET acquisition and result in measured activity to be over-estimated unless subtracted from the initial prompts. The rate of random coincidences on a specific LOR is determined by the equation $R_{i,j} = 2 \cdot t \cdot r_i \cdot r_j$, where $R_{i,j}$ is the random coincidence rate between detector elements i and j , r_i and r_j are singles rates on detector elements i and j , respectively, and t is the coincidence window. So, if the r_i and r_j are measured and t known or determined, the random coincidence rate can be calculated for each LOR. Another way to estimate randoms rate is to use a delayed coincidence time window. In this way the timing signals from one of the coincidence detectors are delayed by more than the timing window resulting in no true coincidences to be detected therefore leaving only random coincidences in the prompts. The latter method has the advantage of having identical dead-time characteristics to the prompt channel but the disadvantage of less statistics as there are significantly less random coincidences than singles detected.

Scatter correction became much more important as the PET technology developed and 2D acquisition was replaced by 3D where the lead septa separating the detector rings were omitted. This, coupled with the increasing transaxial range of the detectors, is contributing to the rapid increase of scattered coincidences detected. For the 3D acquisition mode multiple scatter correction methods have been proposed such as model-based algorithms, „Gaussian fit” technique, multiple energy window, direct measurement, Monte-Carlo modelling and convolution-subtraction techniques. Each of these methods provides specific advantages depending on the scanning geometry and the object scanned. However, there is no universally good method so the scatter correction methods remain active for research.

As considers the technical limitations not all detectors exhibit identical sensitivity and are stable in time. Photomultiplier tubes have variations in gain which may deviate in time. Digital detectors are sensitive to temperature changes and may exhibit variations due to that. Normalisation correction is performed on the system to take into account and offset these differences in between the detectors. Usually, so called direct normalisation is applied, where a coincidence source is used to uniformly expose all detectors at once. In this way all LORs result in a specific number of counts, the inverse of which can then be used as the normalisation coefficient for that LOR in subsequent studies. Since the coefficients can change in time, the normalisation correction calibration needs to be performed on a regular basis.

PET detectors have inherent limits to the rate at which new photons can be detected. This is due to both scintillator afterglow as well as electronics processing rate. The effect can be mitigated when faster electronics and scintillator materials are used. A larger number of discrete detectors would lessen the pile-up of signals on one detector, but the effect cannot be totally eliminated. Losses due to this system dead-time cause the activity in the images to be underestimated at high count rates which needs to be corrected for accurate quantitative results. This is done by modelling the dead-time losses with practical measurements and experiments.

In clinical practice it is essential to perform standardized uptake value (SUV) calibration. For this purpose a uniform phantom filled with known activity contained within an exactly known volume is used. Repeated usually every three months, this procedure assures a correct calibration between the PET image quantitative values provided as counts per second/voxel (cps/voxel) to the actual activity concentration (Bq/ml) within the measured object

1.2.2 Image reconstruction

After the collection and correction of LORs, either analytical or iterative tomographic reconstruction methods are applied to the corrected (projection) data. While both lead to a final 3D dataset, the reconstruction differs a bit depending on whether the data was acquired in 2D or 3D mode. In 2D mode, the LORs covering single axial planes are organised into sets of projections covering 360° . The projections of the imaging planes placed on top of each other as a function of the angle of acquisition form a sinogram. Reconstruction for each of the axial planes is performed separately and the reconstructed slices stacked together form a final 3D image. In 3D acquisition mode, the LORs are available from oblique planes in addition to axial planes, due to the fact that the lead septa, which in 2D mode separated the planes from each other, were removed. This results in an increased scatter contribution as well as a larger dataset that is computationally more demanding.

The reconstruction of an axial PET image from a number of projections can be performed either analytically, which leads to an exact solution but results in poor signal-to-noise ratio as well as resolution properties, or, iteratively, which however computationally more expensive, takes into account the stochastic nature of the data and produces more realistic system model. As considers the analytical 2D image reconstruction, it relies on various methods of back-projecting the acquired projection data into the image matrix which usually includes a filtering step before or after the reprojection in order to counteract the inevitable oversampling in the centre of the Fourier transform and smooth the otherwise noisy image. In 3D analytical reconstruction methods, the data is initially transformed in a form that enables the use of the 2D reconstruction methods, but with the additional statistics provided by the 3D acquisition. In case of iterative reconstruction methods the complexity and accuracy of the system model is increased and takes into account the noise introduced in the acquisition process. This results in a non-deterministic solution to the reconstruction problem, which cannot be solved by analytical means, but, by applying a step-by-step, called iterative approach to the „actual” image. Using iterative reconstruction methods, it is possible to achieve a more accurate estimate of the PET image. Iterative approaches are computationally very expensive and only recently the advances in computational speed and algorithms have allowed the method to be clinically used.

As described, there are multiple ways of reconstructing PET data, most of these being iterative in the modern PET scanners. The most widespread of the reconstruction methods is the ordered subset expectation maximization (OSEM) and its variations. The OSEM method inherently implies that a number of subsets, to which the data is subdivided to, and iteration steps, that are taken during reconstruction, are defined prior to the start of reconstruction. Usually, post-filtering, using Gaussian filtering, is additionally applied to the reconstructed image.

Time-of-flight (ToF) is a method, being available due to faster scintillators and detector electronics, producing more accurate localisation of the annihilation event on the line-of-response, by precisely registering the arrival times of both annihilation photons and calculating the location from the delay between the detection events. ToF leads to a better contrast vs. noise trade-off than non-TOF, especially in case of obese patients [27][28].

To improve PET spatial resolution which is degraded due to scintillator crystal size, penetration and scattering between neighbouring crystals as well as positron range and the photons' deviation from 180° annihilation angle, during the reconstruction process a correction matrix is applied [29][30]. Today, all major vendors of PET/CT systems implement this kind of correction to the data – called point spread function (PSF), resolution recovery or resolution modelling. The correction significantly reduces the spill-out and spill-in effects, caused by incorrect activity estimates due to different activity concentrations in neighbouring regions, and by doing this, the contrast recovery and SUV values are increased. It is especially prominent in small lesions, which become more easily detectable [31][32].

1.2.3 Quantitative imaging in PET

Fundamentally, PET is a quantitative imaging technique and in order to quantitatively image any biomarker, repeatability and reproducibility of measurements is essential. Same results should be achieved not only when a patient is examined repeatedly on the same scanner but also when the exam is performed on another system in another institution and/or from different vendors.

Dynamic acquisition (with arterial blood sampling) is considered to be the „gold standard” in quantitative PET imaging. In this case, data is acquired during multiple time frames over the course of radiopharmaceutical accumulation in the area of interest providing us with information on tracer kinetics. This method, however, is limited due to extensive time required for the acquisition and interpretation. As a simpler alternative, the now widely used semi-quantitative metric, the standardized uptake value (SUV) has been devised. SUV characterizes the activity of a suspect lesion by normalizing its measured activity concentration to injected activity per kg of patient weight. SUV is a widely applied and simple method for PET quantification frequently used in [^{18}F]FDG imaging that allows for a non-invasive tracer kinetics estimate from a static image. SUV, however, is affected by multiple factors (Table 1) – being either technical, biological or physical [33], which can introduce a bias into the data.

The most frequent technical errors are related to incorrect calibration of associated measurement equipment – activity meter and/or PET system – and the synchronisation of PET scanner and activity meter clocks. The effect of these errors can be significantly reduced when adequate quality control regimens are in place. The accuracy and repeatability of activity meter measurements should be regularly verified using appropriate check sources traceable back to the primary standard, while the cross-calibration of PET scanner and activity meter should be performed in accordance with the PET system’s manufacturer’s recommendations. Synchronisation of clocks may nowadays be done automatically on much of the equipment using Network Time Protocol (NTP) servers – this is a convenient way ensuring that the system’s clock is constantly checked against a known value and corrected automatically if required. As setting up an NTP server connection is often not possible on more simple equipment such as the activity meter or if patient injection is carried out manually, a regular verification of the associated clocks by department staff may be necessary. Operator related errors may also occur during every step that requires human intervention. Data entry errors (ex. injection time is used instead of calibration time) are a possible source of variability as long as manual data transfer between different systems remains regular practice. Paravenous administration of activity may occur if the cannula is not placed correctly and the misalignment with the vein is not discovered before the administration of activity. Well trained and experienced staff along with extra checks on critical steps of the process help reduce these kind of operator errors. Critical review of the images and data may retrospectively reveal some of the errors, however

it may not be possible to correct them thereafter, and it helps to avoid misinterpretation of patient data.

Patient biological factors play a major role in quantitative PET imaging. When imaging is done using a tracer that is naturally occurring in the body and the levels of which are fluctuating the effect on the uptake of the tracer can be significant. Uptake of [^{18}F]FDG, being a glucose analogue, is significantly affected by the natural blood glucose levels during the FDG uptake period. If high glucose levels are present during this uptake period, the additionally injected FDG will have to compete with glucose and therefore a smaller amount is accumulated in the tissues, in comparison with a lower glucose level. Glucose measurement prior to administering FDG is a good way to check if the blood glucose levels are not too high. Using a calibrated glucometer, it is possible to take the measured value into account when calculating a glucose corrected SUV. Regardless of radiotracer, if no arterial blood sampling is used, it is difficult to estimate the amount of activity leaving the blood pool and taken up in the organs, which may vary in time and from patient to patient. As a solution the ratio of SUV from two different regions (target and reference) within the same image/study may be calculated, resulting in a SUVR metric. Variations in uptake time of the radiotracer are having impact on quantitative results. For instance, in FDG imaging SUV is known to increase when the time interval between injection and imaging increases. This effect can be offset when a specific uptake time, standard to the imaging procedure is applied. There are additional patient related factors that cannot be totally eliminated such as patient motion during the acquisition – whether due to breathing, peristaltics or comfort during a lengthy procedure or inflammation caused by possibly unrelated and unpredictable conditions.

Acquisition and reconstruction parameters play a major role in achieving accurate and repeatable quantification in PET. Scan duration along with reconstruction parameters such as image matrix size and post filtering directly affects the amount of noise present in the final images. This may in turn affect the quantitative results achieved using some SUV metrics. An upward bias results from using SUVmax on images with higher noise. Reconstruction parameters such as number of iterations and subsets along with post filter values may differ in scanners due to technical limitations, preferences of the reporting physicians and imaging task. This leads to variability in the quantitative results that is dependent on the activity concentration and size of the object imaged as well as the activity in surrounding tissues. Recently developed corrections, such as resolution recovery, present on modern PET/CT systems can significantly increase the SUV of smaller lesions but have a distinct noise texture [31][34] and produce edge overshoot and focal uptake hyper-resolution artefacts [35][36]. Quantification depends on the definition of the measured volume - or volume of interest (VOI) to be measured. In practice a tumour is usually delineated using automatic region drawing methods with cut-off values based on either user defined (i.e. $\text{SUV} \geq 4.0$) or object determined (SUVmax) values. These methods may further vary due to different

implementation by vendors which may include, exclude or use interpolation of voxel values for lesion border definition. As there are many factors affecting the reconstruction and measurement, which cannot be precisely characterised beforehand, it is useful to practice harmonisation of the resulting recovery coefficients.

Factor	Mitigation
Technical errors	
Activity meter calibration error	Regular quality checks and traceable calibration
Cross-calibration of activity meter and PET	PET is regularly calibrated using same activity meter, harmonisation programs
Incorrect synchronisation of clocks on PET system and activity meter	Regular checks by staff or automated synchronisation if possible
Residual activity after administration	Residual activity is measured and taken into account
Data entry errors	Use of automated transfer of data
Paravenous administration of activity	Test bolus injection, retrospective image analysis
Biologic factors	
Blood glucose level	Specific fasting and measurement of blood glucose before the study
FDG clearance from blood	Normalise SUV to blood uptake (SUV _r)
Uptake period	Standardised uptake period
Patient motion	Comfortable position, gating systems, correction algorithms
Inflammation	Pre-scan patient questionnaires
Physical factors	
Acquisition parameters	Noise reduction algorithms, adapt protocol for
Reconstruction parameters	Harmonisation programs
Region of interest (ROI) used	Automated standardised tools for repeatable lesion delineation
SUV type used	Clear specification of used metrics
Blood glucose level correction	Proper fasting procedures
Use of contrast agents during CT-AC	Scanning in venous phase, specific CT corrections

Table 1. Factors affecting [¹⁸F]FDG PET quantification.

1.3 Multicentre standardisation programs

Due to the wide spread of PET/CT systems, the number of multicentre trials (such as ABLE, RAPID, FALCON and many others) involving them has increased throughout recent years, and quantitative evaluation and endpoints are more frequently used [37][38]. In order to achieve best multicentre standardisation and harmonisation, the sources of variability should be considered and their effects minimised as much as possible. Rapid advancement and different levels of technology in various imaging centres pose a particular problem to standardisation as the benefits offered by modern technologies, such as PSF or ToF, provide increased benefit over the older technologies. This creates difficulties in comparing the data acquired on different generation systems. There is a number of guidelines and standardisation programs developed by professional and scientific societies and organisations which aim at reducing this variability by standardising the involved procedures and/or harmonising the PET/CT system performances [39–43].

Society of Nuclear Medicine and Molecular Imaging (SNMMI) has been running the Clinical Trials Network (CTN) since 2008 to ensure baseline common quality control metrics for participating PET scanners used in multi-centre studies. Scanner validation requires the use of program specific CTN phantom – an anthropomorphic phantom mimicking a human torso to provide qualitative and quantitative information in conditions similar to clinical imaging settings. An acceptance criteria of $\pm 10\%$ is set for the background region SUVmean, but precise sphere-specific acceptance criteria for SUVmax and SUVmean for the various sphere sizes are currently not strictly set [44]. Scanner re-validation, including phantom scanning, is repeated annually.

From 2003, the American College of Radiology Imaging Network (ACRIN) PET Core Laboratory qualifies sites for participation in multicentre trials. To do this, each site is required to submit data from one uniform cylinder scan and two patient test cases. The uniform cylinder scan is analysed in the core lab by calculating the mean SUV of each transverse slice, while in order to pass, the average SUV is required to be within 1.0 ± 0.1 . Patient scans are qualitatively evaluated for artefacts, noise, patient positioning and PET - CT image alignment. In addition, the core lab calculates SUVs from the ROIs comprising the livers of submitted patients' scans and compares them with the reported data. This is aimed at evaluating the level of expertise of PET/CT camera operators. No acceptance criteria for size specific contrast recovery coefficients is set [45].

Radiological Society of North America - Quantitative Imaging Biomarkers Alliance (RSNA-QIBA) provides guidelines and instructions on how to measure SUV calibration as well as uniformity between slices of a uniform cylindrical phantom. „Acceptable” and „ideal”

performance measurement methods and criteria are described for uniformity, resolution/SUV recovery and noise. Criteria of ACRIN/EANM/SNM are frequently referred to. The results of the quality control measurements are to be submitted annually and should be available for any site audit [46].

National Cancer Research Institute (NCRI) PET Clinical Trials Network provides more a specific accreditation on a per trial basis, where the Core Lab reviews the trial protocol in the setup phase and determines specific requirements for the participating system's quality assurance. This results in a uniform distribution of scanners within a single trial, but, requires multiple levels of standardisation which creates difficulties in intercomparison.

The European Association of Nuclear Medicine (EANM) published the guideline „FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0” in 2010 [47] which was followed by and updated version 2.0 in 2014 [48]. As a support for the clinical application of the EANM guideline, the EANM-EARL initiative provides an accreditation service based on regular independent review of phantom scans which are required to be performed by the imaging sites participating in the program. Quarterly, a uniform cylinder needs to be scanned and data submitted to be evaluated centrally for SUV calibration check in order to ensure that the mean SUV is within the acceptance range of 1.0 ± 0.1 . Annually, the NEMA body phantom should be scanned for the evaluation of image quality and contrast recovery of the six hot spheres of varying diameters (10 – 37 mm). In order that the imaging sites maintain their accredited status, the contrast recoveries mean and max SUVs needs to comply with the accreditation specifications. In addition to periodically providing phantom scans, the accredited sites have to apply the EANM procedure guidelines [48].

EANM EARL aims at constantly adapting, improving and expanding based on recent scientific discoveries, technical developments of the equipment and requirements of the nuclear medicine community. Work described in this thesis focuses on the EANM guidelines and updating/extending the associated EARL accreditation program for quantitative multicentre PET/CT imaging.

1.4 Thesis aim

EANM-EARL [^{18}F]FDG PET/CT accreditation programme has been running successfully from 2010, applying the standard criteria developed by EARL for the scanners used in clinical practices. As new technology appears on the market and advances are being made in using novel tracers, the existing quantitative imaging standardisation programs need to be adapted and evolve along with the field, in order to remain relevant in the constantly changing environment

This thesis aims at investigating the current status and results of the EANM-EARL [^{18}F]FDG PET/CT accreditation programme and explore the feasibility of an update on the current harmonising standards.

Another goal of this work is to expand the EANM-EARL accreditation program to include additional radioisotopes and tracers besides [^{18}F]FDG.

1.5 Thesis outline

Chapter 2 describes six years' experience of running the EANM-EARL [^{18}F]FDG PET/CT accreditation programme and reports the findings and impact on the participating nuclear medicine centres. In **Chapter 3** the feasibility of harmonising state-of-the-art PET/CT systems equipped with time-of-flight (ToF) and resolution modelling capabilities is explored. Additionally, a prototype of updated harmonising criteria is developed and suggested for application by EARL. **Chapter 4** investigates the impact of updated EARL harmonising criteria on quantitative reads of clinical PET/CT studies and tests a method of bridging the old and new standards. In **Chapter 5** the feasibility of harmonisation in quantitative ^{89}Zr imaging is investigated.

Bibliography

1. Röntgen WC. On a new kind of rays. *Science* (80-). 1896;
2. Becquerel AH. Sur les radiations invisibles émises par les corps phosphorescents. *CR Acad. Sci. Paris*. 1896;
3. Squires G. J J Thomson and the discovery of the electron. *Phys. World* [Internet]. 1997;10:33–6. Available from: <http://stacks.iop.org/20587058-/10/i=4/a=25?key=crossref.5d87e63fe4a1f7a0b3c04c79794b207a>
4. Flakus FN. Radiation detection Detecting and measuring ionizing radiation - a short history. *IEAE Bull*. 1981;
5. Kallmann H. Quantitative measurements with scintillation counters. *Phys. Rev*. 1949;
6. Cassen B, Reed CW, Curtis L, University of California LA, Commission. USAE. A sensitive directional gamma ray detector. Oak Ridge, Tennessee: United States Atomic Energy Commission, Technical Information Service; 1949. p. 7 pages.
7. Wrenn FR, Good ML, Handler P. The use of positron-emitting radioisotopes for the localization of brain tumors. *Science* (80-). 1951;
8. Brownell GL, Sweet WH. Localization of brain tumors with positron emitters. *Nucleonics*. 1953;
9. Rankowitz S, Robertson JS, Higinbotham WA, Rosenblum MJ. Positron scanner for locating brain tumors. *IRE Int. Conv. Rec*. 1961.
10. O Anger H, Gottschalk A. Localization of brain tumors with the positronscintillation camera. *J. Nucl. Med*. 1963.
11. Hounsfield GN. Computerized transverse axial scanning (tomography): I. Description of system. *Br. J. Radiol*. 1973;
12. Ambrose J. Computerized transverse axial scanning (tomography): II. Clinical application. *Br. J. Radiol*. 1973;
13. Ambrose JA. Computerized transverse axial tomography. *Br. J. Radiol*. 1973;
14. Phelps ME, Hoffman EJ, Mullani NA, Ter-Pogossian MM. Application of annihilation coincidence detection to transaxial reconstruction tomography. *J. Nucl. Med*. 1975;
15. Phelps ME, Hoffman EJ, Mullani NA, Higgins CS, Pogossian MMT. Design considerations for a positron emission transaxial tomograph (Pett III). *IEEE Trans. Nucl. Sci*. 1976;
16. Cho ZH, Farukhi MR. Bismuth Germanate as a Potential Scintillation Detector in Positron Cameras. *J. Nucl. Med*. 1977;
17. Mullani NA, Gaeta J, Yerian K, Wong WH, Hartz RK, Philippe EA, et al. Dynamic imaging with high resolution time-of-flight pet camera-TOFPET I. *IEEE Trans. Nucl. Sci*. 1984;
18. Rich DA. A Brief History of Positron Emission Tomography. *J. Nucl. Med. Technol.* . 1997;25:4–11.
19. Zhang J, Maniawski P, Knopp M V. Performance evaluation of the next generation solid-state digital photon counting PET/CT system. *EJNMMI Res*. 2018;
20. Lopci E, Nanni C, Castellucci P, Montini GC, Allegri V, Rubello D, et al. Imaging with non-FDG PET tracers: outlook for current clinical applications. *Insights Imaging*. 2010;
21. Narayanaswami V, Dahl K, Bernard-Gauthier V, Josephson L, Cumming P, Vasdev N. Emerging PET Radiotracers and Targets for Imaging of Neuroinflammation in Neurodegenerative Diseases: Outlook Beyond TSPO. *Mol. Imaging*. 2018.
22. Davidson CQ, Phenix CP, Tai T, Khaper N, Lees SJ. Searching for novel PET radiotracers: imaging cardiac perfusion, metabolism and inflammation. *Am J Nucl Med Mol Imaging*. 2018;
23. Hofmann M, Steinke F, Scheel V, Charpiat G, Farquhar J, Aschoff P, et al. MRI-Based Attenuation Correction for PET/MRI: A Novel Approach Combining Pattern Recognition and Atlas Registration. *J. Nucl. Med*. 2008;
24. Hofmann M, Steinke F, Scheel V, Charpiat G, Brady M, Schölkopf B, et al. MR-Based PET Attenuation Correction: Method and Validation. 2007 IEEE Nucl. Sci. Symp. Med. Imaging Conf. (NSS-MIC 2007). 2007.
25. Olin A, Ladefoged CN, Langer NH, Keller SH, Löfgren J, Hansen AE, et al. Reproducibility of MR-Based Attenuation Maps in PET/MRI and the Impact on PET Quantification in Lung Cancer. *J. Nucl. Med*. 2018;
26. Kuttner S, Lassen ML, Øen SK, Sundset R, Beyer T, Eikenes L. Quantitative PET/MR imaging of lung cancer in the presence of artifacts in the MR-based attenuation correction maps. *Acta radiol*. 2019;
27. Karp JS, Surti S, Daube-Witherspoon ME, Muehllehner G. Benefit of Time-of-Flight in PET: Experimental and Clinical Results. *J. Nucl. Med*. 2008;
28. Surti S, Kuhn A, Werner ME, Perkins AE, Kolthammer J, Karp JS. Performance of Philips

- Gemini TF PET/CT scanner with special consideration for its time-of-flight imaging capabilities. *J. Nucl. Med.* 2007;
29. Sureau FC, Reader AJ, Comtat C, Leroy C, Ribeiro M-J, Buvat I, et al. Impact of Image-Space Resolution Modeling for Studies with the High-Resolution Research Tomograph. *J. Nucl. Med.* 2008;
 30. Panin V, Kehren F, Michel C, Casey M. Fully 3-D PET Reconstruction with system matrix derived from point source measurements. *IEEE Trans Med Imaging* [Internet]. 2007;25. Available from: <http://dx.doi.org/10.1109/TMI.2006.876171>
 31. Tong S, Alessio AM, Kinahan PE. Noise and signal properties in PSF-based fully 3D PET image reconstruction: An experimental evaluation. *Phys. Med. Biol.* 2010;
 32. Andersen FL, Klausen TL, Loft A, Beyer T, Holm S. Clinical evaluation of PET image reconstruction using a spatial resolution model. *Eur. J. Radiol.* 2013;
 33. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J. Nucl. Med.* 2009;50 Suppl 1:11S-20S.
 34. Rahmim A, Tang J. Noise propagation in resolution modeled PET imaging and its impact on detectability. *Phys. Med. Biol.* 2013;
 35. Rahmim A, Qi J, Sossi V. Resolution modeling in PET imaging: Theory, practice, benefits, and pitfalls. *Med. Phys.* 2013;
 36. Watson CC. Estimating effective model kernel widths for PSF reconstruction in PET. *IEEE Nucl. Sci. Symp. Conf. Rec.* 2012.
 37. Doot RK, Kurland BF, Kinahan PE, Mankoff DA. Design Considerations for using PET as a Response Measure in Single Site and Multicenter Clinical Trials. *Acad. Radiol.* 2012;
 38. Yankeelov TE, Mankoff DA, Schwartz LH, Lieberman FS, Buatti JM, Mountz JM, et al. Quantitative imaging in cancer clinical trials. *Clin. Cancer Res.* 2016.
 39. SNMMI Clinical Trials Network (CTN) [Internet]. Available from: <http://www.snmmi.org/Research/Content.aspx?ItemNumber=9937&navItemNumber=6832>
 40. American College of Radiology Imaging Network (ACRIN) [Internet]. Available from: <https://www.acrin.org/CORELABS.aspx>
 41. Radiological Society of North America - Quantitative Imaging Biomarkers Alliance (RSNA-QIBA) [Internet]. Available from: <https://www.rsna.org/en/research/quantitative-imaging-biomarkers-alliance>
 42. UK PET Core Lab [Internet]. Available from: <http://www.ncri-pet.org.uk/>
 43. EANM EARL website [Internet]. Available from: <http://earl.eanm.org>
 44. Sunderland JJ, Christian PE. Quantitative PET/CT Scanner Performance Characterization Based Upon the Society of Nuclear Medicine and Molecular Imaging Clinical Trials Network Oncology Clinical Simulator Phantom. *J Nucl Med.* 2015;56:145–52.
 45. Scheuermann JS, Saffer JR, Karp JS, Levering AM, Siegel A. Qualification of PET Scanners for Use in Multicenter Cancer Clinical Trials: The American College of Radiology Imaging Network Experience. *J. Nucl. Med.* 2010;50:1187–93.
 46. Kinahan P, Wahl R, Shao L, Frank R, Perlman E. The QIBA profile for quantitative FDG-PET/CT oncology imaging. *J. Nucl. Med.* 2014;55:1520–.
 47. Boellaard R. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 1.0. *Eur. J. Nucl. Med. Mol. Imaging.* 2009;
 48. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur. J. Nucl. Med. Mol. Imaging.* 2014;42:328–54.

Chapter 2

EANM/EARL FDG-PET/CT accreditation - summary results from the first 200 accredited imaging systems

Author(s): Andres Kaalep, Terez Sera, Win Oyen, Bernd J. Krause,
Arturo Chiti, Yan Liu, Ronald Boellard

*as published in the European Journal of Nuclear Medicine and Molecular Imaging
(Kaalep et al, 2017)*

2.1 Abstract

Purpose: From 2010 until July 2016 EANM Research Ltd. (EARL) FDG-PET/CT accreditation program has collected over 2500 phantom datasets from approximately 200 systems and 150 imaging sites worldwide. The objective of this study is to report the findings and impact of the accreditation program on the participating PET/CT systems.

Methods: To obtain and maintain EARL accredited status, sites were required to complete and submit two phantom scans - calibration quality control (CalQC), using a uniform cylindrical phantom and image quality control (IQQC), using a NEMA NU2-2007 body phantom. Average volumetric SUV bias and SUV recovery coefficients (RC) were calculated and the data evaluated on the basis of quality control (QC) type, approval status, PET/CT system manufacturer and submission order.

Results: SUV bias in 5% ($n=96$) of all CalQC submissions ($n=1816$) exceeded 10%. After corrective actions following EARL feedback, sites achieved 100% compliance within EARL specifications. 30% ($n=1381$) of SUVmean and 23% ($n=1095$) of SUVmax sphere recoveries from IQQC submissions failed to meet EARL accreditation criteria while after accreditation, failure rate decreased to 12% ($n=360$) and 9% ($n=254$) respectively. Most systems demonstrated longitudinal SUV bias reproducibility within $\pm 5\%$, while RC values remained stable and generally within $\pm 10\%$ for the four largest and $\pm 20\%$ for the two smallest spheres.

Conclusions: Regardless of manufacturer or model, all investigated systems are able to comply with the EARL specifications. Within the EARL accreditation program gross PET/CT calibration errors are successfully identified and longitudinal variability in PET/CT performances reduced. The program demonstrates that a harmonising accreditation procedure is feasible and achievable.

2.2 Introduction

Positron emission tomography (PET) and computed tomography (CT) hybrid imaging (PET/CT) using ^{18}F -fluorodeoxyglucose (FDG) has become a routinely used and valuable tool in oncology. It is widely utilised for diagnosis, staging and restaging of various malignancies [1–12] as well as response monitoring due to its ability to measure metabolic changes [13–19]. Standard uptake value (SUV), which represents the tissue radioactivity concentration normalized to injected activity and body weight [20] is the most frequently used quantitative metric in oncology [21,22]. Multiple factors, however, can give rise to bias [23–25] and increased variability in SUV, especially when inter-centre comparison is required from institutions lacking a uniform approach to imaging procedures [26–28]. The variability is a significant issue for clinical trials or multicentre studies utilising the quantitative potential of PET [24,26–31]. In clinical practice, there is a wide range of PET systems installed globally including scanners developed more than 10 years ago along with brand new devices incorporating state of the art acquisition (i.e. time of flight, digital PET detectors) and reconstruction (i.e. resolution modelling) technologies [32]. In addition to various PET/CT models available, the acquisition and reconstruction parameters applied at different sites vary greatly due to local preferences [24,32,33]. Centres equipped with PET systems having new acquisition and reconstruction technologies available, often tend to aim for the possible best lesion detection which may not be in line with quantitative harmonising standards [34]. Aforementioned technical factors impose a significant source of variability in PET quantification [24,32] that should be addressed by the international community.

Numerous professional societies and organizations such as the Society of Nuclear Medicine and Molecular Imaging (SNMMI), American College of Radiology Imaging Network (ACRIN), Radiological Society of North America - Quantitative Imaging Biomarkers Alliance (RSNA-QIBA), The American Association of Physicists in Medicine (AAPM) and the European Association of Nuclear Medicine (EANM) are promoting harmonisation of imaging procedures [35–37] in order to reduce the variability of PET image quantification in a multicentre setting. Many of these programs rely on quality control procedures utilising standard phantoms [38] for standardization of quantification [32,39–41] and harmonisation of PET/CT systems [35]. Review papers on describing some of the results and experience in running such programs have been published by Scheuermann et al. [39] and more recently by Sunderland et al. [32].

In 2006, the European Association of Nuclear Medicine (EANM) launched the EANM Research Ltd (EARL) initiative. One of the main objectives of the program has been promoting multicentre nuclear medicine and research. In 2010, the FDG-PET/CT accreditation program was created in order to address variability in the quickly growing field of quantitative FDG-PET imaging by setting up guidelines and specifications to which the participating sites must adhere to.

The specification bandwidths for the current EARL specifications were developed during a pilot study in 2010-2011 involving 12 PET/CT systems. Based on this study, specifications for SUVmean and SUVmax recovery coefficients were derived which accommodated all investigated systems. From its initiation until July 2016, EARL has collected approximately 2500 phantom datasets from more than 200 PET/CT systems from over 150 imaging sites worldwide. The data analysed by EARL encompasses majority of the system types in clinical use over the past 10 years and incorporates sites with various backgrounds giving it a broad basis to represent the field as a whole.

The objective of this paper is to report the findings obtained so far in the EARL standardisation program and their impact on the quantitative variability of accredited PET/CT systems. Analysis of phantom scans from the largest number of active PET centres so far provides representative details of current quantitative capabilities of FDG-PET imaging and the variability to be expected. Understanding the characteristics of variability and the impact on standardization will help planning multi-centre clinical trials, utilising quantitative FDG-PET/CT imaging and advance use of PET as a quantitative imaging biomarker.

The secondary objective of this study is to explore ways to improve the EARL FDG-PET accreditation program based on the retrospective analysis of phantom data collected in the EARL database.

2.3 Materials and methods

2.3.1 Acquisition and submission of data to EARL

Sites, which are seeking EARL FDG-PET/CT accreditation for the first time, need to pass the initial procedure. This procedure includes the submission of an online questionnaire and a signed statement – these documents have to be submitted at the start of the accreditation procedure and revised annually, whereas QC documents need to be regularly provided in order to maintain the EARL accredited status.

For the first and follow up procedures, sites have to perform calibration QC and image quality QC measurements. The calibration QC measurements have to be repeated every 3 months and image quality QC procedures annually, while the data needs to be provided to EARL upon completion of the procedures. During each round of QC survey there is a 3 week period for the sites to collect the data and submit it to EARL, followed by a 3 week period of analysing the data by EARL and reporting the results back to the sites.

For the calibration QC measurements, centres are asked to use a cylindrical phantom with the following characteristics: diameter of about 20 cm (17 to 22 cm) and length sufficient to cover the entire axial field of view (FOV). Furthermore, the exact volume of the calibration phantom should be known and recorded in the calibration QC scan report form. The phantom has to be filled with water and about 70 MBq ^{18}F -FDG added to it, aimed at expected phantom acquisition time.

For image quality QC measurements, the NEMA NU2-2007 image quality phantom is required. The phantom has a fillable torso cavity to act as a background compartment, a 5 cm diameter cylindrical lung insert in the centre and 6 fillable spheres with internal diameters of 10 mm, 13 mm, 17 mm, 22 mm, 28 mm and 37 mm positioned coaxially around the lung insert. The phantom background compartment and the spherical inserts have to be filled with ^{18}F -FDG solution aimed at activity concentrations at the start of the PET scan of 2 kBq/mL and 20 kBq/mL, respectively, resulting in a sphere to background ratio of 10:1.

With both phantoms, routine quantitative whole body PET/CT scans have to be performed with 2 PET bed positions of at least 5 min each, including a (low dose) CT for attenuation correction purposes [35]. After reconstruction, the attenuation corrected PET, non-attenuation corrected PET and CT images of the phantoms have to be uploaded into the EARL central database, along with scan report forms.

2.3.2 Quantitative analysis and approval by EARL

The uniform calibration QC phantom and NEMA NU2 IQ body phantom images uploaded into the EARL database are evaluated centrally, making use of a standardized semi-automatic quantitative analysis tool developed internally within EARL. The software uses activity and time information provided by the scan report forms. The average volumetric SUV bias is generated as relative deviation between measured and calculated activity concentration values (equation 1). The SUV recovery coefficients (RCs) for the 6 spherical inserts are based on 50% background corrected isocontour VOI ($\text{RC}_{\text{SUVmean}}$) and maximum voxel value included in the VOI ($\text{RC}_{\text{SUVmax}}$).

$$\text{SUVbias (\%)} = \left(\frac{C_{\text{measured}}}{C_{\text{calculated}}} - 1 \right) \times 100\%; \text{ (Eq.1)}$$

,where

C_{measured} - activity concentration measured from images

$C_{\text{calculated}}$ - activity concentration calculated from injection data

EARL is applying SUV bias and RC values acceptance criteria, which were defined by feasibility studies performed on the systems used in clinical practices at the start of the standardisation - a study is underway in order to update these. When approval is not granted, the site undergoing (re-)accreditation is asked to take corrective actions, for example: recalibration of the PET system, adjustment of reconstruction parameters, repeating the phantom scan and so on. When required, EARL is advising the sites. A Manual describing the accreditation program in detail as well as information on the EARL website [42] is also available. If submitted QC documents meet the standard requirements, FDG-PET/CT accreditation is granted and the department is listed on the EARL website (<http://earl.eanm.org>) as an accredited PET/CT centre of excellence. Furthermore, the site is provided with an accreditation certificate and signet, which can be used on its correspondence and website.

2.3.3 Data clean-up and preparation

To allow for data extraction, the EARL database had to be cleaned of duplicates and entries with insufficient or missing information removed, entry errors were identified and the individual site identification data ignored thereby providing an anonymised set of data for evaluation. First and subsequent site submissions were identified and marked as such.

2.3.4 Analysis

The calibration QC and image quality QC datasets from the EARL database will be analysed based on the type of the phantom, accreditation approval status, manufacturer of the PET/CT system and whether it was the first or a subsequent QC data submission. The SUV bias and normalised SUV biases were analysed as well as the recovery coefficients for each sphere size, separately for SUVmean and SUVmax. For each parameter, mean, median, standard deviation, standard error and skewness were calculated. Longitudinal reproducibility analysis was performed on 16 systems (systems A to P) selected based on each having sufficient longitudinal data of at least 18 approved CalQC data submissions or at least 5 approved IQQC datasets.

2.4 Results

2.4.1 General overview

Data reviewed in this paper encompasses all submissions to the EARL database from the initiation of the standardisation program in November 2010 to July 2016. Figure 1 represents the number of sites and systems participating each year. After correcting for erroneous, partial and duplicate entries, 1816 CalQC and 778 IQQC datasets were used for further analysis. The datasets were 29% (n=752) from GE-, 29% (n=741) from Philips- and 42% (n=1101) from Siemens-systems.

First data submissions constitute 10% (n=175) of all CalQC and 23% (n=178) of all IQQC scans. 85% (n=149) of the first and 94% (n=1537) of subsequent CalQC data submissions could be approved by EARL. This results in an overall approval rate for CalQC of 93% (n=1686). Table 1 states descriptive statistics for CalQC initial and subsequent submissions.

Out of all systems (n=200) that have enrolled in the program, the accreditation for 47 systems (24%) has been discontinued for various reasons, such as scanner replacement or stopped participation in trials requiring EARL accreditation.

2.4.2 Calibration QC

Detailed descriptive statistics for CalQC SUV bias are summarised in Table 1. Figure 2 demonstrates CalQC SUV bias distribution for all, initial and subsequent submissions along with vendor based distribution of approved results. 3% (n=60) of all CalQC submissions were below and 2% (n=36) above the corresponding EARL SUV bias limits of -10% and +10%. 9% (n=16) of systems could not be approved at first CalQC submission, but after corrective actions all of the scanners fulfilled the EARL specifications. Significant mean SUV biases of -1.53% ($p < 0.0001$) and -1.78% ($p < 0.0001$) were observed in approved datasets from GE and Philips systems respectively, while datasets from Siemens systems did not demonstrate this deviation. In figure 3 longitudinal CalQC volumetric SUV bias is plotted as a function of the order of subsequent submissions.

Fig 1. Number of sites and PET/CT systems participating in the EARL accreditation program. For 2016 data has been collected from January to July.

CalQC	Mean SUV bias (%)	Median SUV bias (%)	SUV bias Std. Dev (%)	Skewness	Submissions with SUV bias below EARL specs	Submissions with SUV bias above EARL specs	Submissions with SUV bias within EARL specs
All CalQC	-1.14 (± 0.13)	-1.01	5.36	-0.32	3%	2%	95%
All approved CalQC	-0.97 (± 0.09)	-0.94	3.71	0.15	0%	0%	100%
Sites' first submitted CalQC	-1.25 (± 0.46)	-0.79	6.06	-1.14	6%	3%	91%
Subsequent approved submissions CalQC	-1.01 (± 0.09)	-1.02	3.66	0.16	0%	0%	100%
All approved GE CalQC	-1.53 (± 0.15)	-1.60	3.27	0.31	0%	0%	100%
All approved Philips CalQC	-1.78 (± 0.18)	-1.71	3.89	0.26	0%	0%	100%
All approved Siemens CalQC	-0.05 (± 0.14)	0.06	3.65	0.01	0%	0%	100%

Table 1. CalQC SUV bias statistics from first, regular ongoing and all EARL approved submissions (pooled and per vendor).

Fig 2. Comparison of CalQC SUV bias distribution for all, initial and subsequent submissions along with vendor based distribution of approved results. The dotted horizontal lines represent EARL specification limits. Central line of the box is the median, edges of the box are the 25th and 75th percentiles, the whiskers extend to either the most extreme data points which are not considered outliers or 1.5 times interquartile range. The outliers are marked using plus signs.

Fig 3. Longitudinal plots of EARL approved CalQC results from the 16 systems. SUV bias deviation from the expected value. The dotted horizontal lines represent EARL specification limits. Central line of the box is the median, edges of the box are the 25th and 75th percentiles, the whiskers extend to either the most extreme data points which are not considered outliers or 1.5 times interquartile range. The outliers are marked using plus signs.

2.4.3 Image Quality QC

Descriptive statistics for IQQC SUVmax and IQQC SUVmean results for each sphere size are summarised in supplemental tables 1 and 2. Supplemental figure 4 shows the distribution of sphere recoveries of all submitted data and approved IQQC submissions along with SUV recovery distributions separately per manufacturer. A large variability in sphere recoveries was observed. Out of all sphere recoveries from IQQC datasets, 30% (n=1381) of SUVmean and 23% (n=1095) of SUVmax recoveries failed to meet the EARL accreditation criteria. After corrective actions the corresponding values dropped to 12% (n=360) and 9% (n=254) respectively. A positive bias in the recovery values was observed prior to and also after the corrective actions and granted accreditation, respectively. 47% (n=84) of sites' initial IQQC submissions and 68% (n=409) of all consecutive IQQC submissions were approved by EARL. Overall approval rate for IQQC submissions was 63% (n=493).

To evaluate the longitudinal performance of the 16 systems, sphere recoveries for all sphere sizes were plotted based on the order of subsequent submissions, as seen in supplemental figure 5. From the figure, it can be seen that for each sphere there is an initial large variability in observed SUV recoveries which is reduced and becomes harmonized during subsequent submissions.

2.5 Discussion

The number of sites participating in the EARL accreditation program is steadily increasing and the received data is almost equally distributed among the three major PET/CT system manufacturers.

Inconsistency was observed in the names provided for same types of systems and in some cases the device serial numbers were missing or had been changed at the occasion of software upgrade or service maintenance. This complicated the distinction between new systems and those already existing in the EARL database. As a solution, the EARL database client portal could be configured so that the system type selection be performed by choosing from a predefined list, in the same way as it currently works for system manufacturer, permitting that the regularly submitted QC data be checked and if necessary, corrected for constancy of the core data and the device serial number in particular.

A small number ($n=19$) of calibration QC submissions were not included in further analysis due to large SUV biases observed. The high values might have been related to improper data communication from the participants, inhomogeneous phantoms or system failures left unnoticed by the centres. In all these cases, sites were asked to implement corrective actions and redo the experiments. As a result, all of the affected systems achieved compliance with the EARL specifications and were granted accreditation.

Regardless of manufacturer or model, all systems were able to comply with the CalQC specifications set up by EARL. Only 9% of sites first QC submissions and 5% of all CalQC submissions demonstrated a measured activity bias of more than 10%. Scheuermann et al. in their review of ACRIN PET Core Laboratory program reported a similar initial SUV or normalization calibration failure rate of 12% within the same acceptance criteria of $SUV\ 1.0 \pm 0.1$ [39].

Compared to all received data, the non-compliances were almost eliminated in EARL approved data. In the datasets reviewed and approved by EARL, all CalQC SUV biases fall within the range of $\pm 10.5\%$, which aligns with the target of $\pm 10\%$. Although the fraction of non-approved data from combined first and regular submissions was relatively small (about 5%), the importance of this fact should not be underestimated since noncompliance in calibration QC procedure means gross error in basic system calibration which would affect all further PET quantitative evaluations.

CalQC average values demonstrate a slight but statistically significant underestimation of the activity concentration or SUV by Philips and GE systems, while data from Siemens devices do not show this deviation. Scheuermann et al. reported similar results for Philips systems [39]. Whether this is due to some systematic differences among the vendors' calibration procedures, drifts in calibration values or some unknown effects, could not be derived from the current data but should be subject for further investigations. However, it is important to note that these errors were well within 2% and are very likely to be clinically irrelevant.

Longitudinal analysis of the CalQC results from 16 selected systems, visualised in supplemental figure 6, illustrates good performance of the systems and reflects the quantification stability achieved by the accredited sites. Large SUV biases were only occasionally observed, and were all corrected by the sites after receiving notification from EARL. In most of the cases corrective action had been taken within 4 weeks.

The majority of PET/CT systems followed throughout the investigation period showed good reproducibility of the CalQC results seen in figure 3. Longitudinal SUV bias values lie mostly within $\pm 5\%$ of the expected value, which is in agreement with data published by Geworski et al. [43] and more recently by Sunderland et al. [32]. The rest of the systems also meet the $\pm 10\%$ calibration accuracy criteria. Our findings suggest opportunities for the implementation of stricter accreditation specification for EARL CalQC SUV biases.

The comparison between all received and EARL approved data indicates significant reduction of outliers for IQQC results (supplemental tables 1 and 2). However, some of the data not strictly within EARL specifications was still accepted after critical review. This means the noncompliant data was deemed acceptable in case only one of the spheres being slightly out of the specifications, but the overall shape and magnitude of RC curve could be determined to a sufficient extent and being compliant with EARL criteria. Moreover, in these cases submissions were reviewed retrospectively and consistency of acquisition, reconstruction and settings was verified. While in case of the larger diameter spheres meeting EARL specifications was common, for the smallest sphere (10 mm diameter) SUVmax data remained outside of specified bandwidth in 26% of the EARL approved submissions. Achieving harmonised quantitative results for the smallest sphere turned out to be challenging due to its increased sensitivity to image noise and phantom positioning within the PET field of view.

Although in most cases when the spheres' RCs did not meet the specifications, sites were asked to redo experiments, still a noticeable upward bias in results for the smallest sphere could be observed. Therefore, based on these findings EARL implemented slightly adjusted recovery specifications for the smallest sphere starting from the 1st of January 2017.

Prior to harmonisation, the average recoveries of all spheres demonstrated a positive bias compared to EARL specifications. The trend could still be observed to some extent within the approved results. This can be attributed to both the vendors and sites preferences leaning towards reconstruction algorithms emphasizing contrast and detectability. The positive bias in the case of 10 mm diameter sphere might also be attributed to newly adopted acquisition and reconstruction technologies (e.g. including resolution modelling), which aims at increasing small lesion detectability. This trend was also observed by Sunderland et al. who showed that recoveries were generally higher for systems employing TOF and PSF reconstructions [32]. Since more modern PET/CT systems with new technologies appear in the field, a review of the existing EARL specifications is required in order to facilitate the inclusion of the increased contrast recovery capabilities of these systems. EARL is currently carrying out a feasibility study aiming at redefining the accreditation criteria by taking into account not only the new technologies but also considering that the majority of the PET/CT systems currently in clinical use should be able to comply.

By comparing the first and following regular submissions from participating sites, a relative increase can be observed in meeting the EARL specifications which is expected since the accredited sites gain experience in performing quantitative calibrations and assessment of their systems. The staff's increased awareness towards the necessity of regular calibration and quality control of the systems is one of the benefits of participation in an accreditation/standardisation program, which is difficult to measure but should not be underestimated.

IQQC longitudinal analysis results, presented in figure 5, visualize the major improvement of regular submissions in relation to the first submission. As the sphere size decreases, the improvement becomes more prominent. After initial adjustment of the reconstruction parameters, the longitudinal reproducibility remains stable and generally within $\pm 10\%$ for the largest 4 spheres and $\pm 20\%$ for the smallest 2 spheres. These findings emphasise that PET/CT performance and the execution of the QC experiments show high reproducibility and demonstrate that long term maintenance of a harmonised PET/CT network is feasible and achievable.

2.6 Conclusion

The European Association of Nuclear Medicine (EANM) has been running an FDG-PET/CT accreditation program under the EANM Research Ltd (EARL) initiative in order to harmonise quantitative PET/CT performance and facilitate multi-centre nuclear medicine and research. The number of sites and systems participating in the accreditation program has steadily increased over the years with similar numbers of scanners from each of the 3 major vendors.

Outliers observed in the overall submissions of both CalQC and IQQC were largely eliminated in subsequent submissions after notification from EARL. Excellent longitudinal performance was observed in most of the systems – majority demonstrated CalQC values reproducible within 5% and IQQC results within 10% for the largest 4 spheres and 20% for the smallest two spheres. Occasional deviations from expected values were rapidly resolved by the sites after notification from EARL. Regardless of system manufacturer or model, all vendors were able to comply with the accreditation specifications set out by EARL.

Prior to harmonisation, IQQC data demonstrated a slight positive bias relative to EARL specifications, which suggested to carry out a review and update in order to account for the advances in acquisition and reconstruction technologies in PET/CT.

In this manuscript we have demonstrated that the EARL accreditation program can successfully identify gross PET/CT calibration errors and reduces variability in PET/CT performance by longitudinally performing harmonisation QC experiments. The program is running successfully for more than 6 years and shows that a harmonising accreditation procedure is feasible and achievable.

Centres with accredited PET/CT systems benefit greatly from participating in large-scale accreditation programs which facilitate the implementation of procedural guidelines widely recognised by the imaging community.

2.7 Compliance with Ethical Standards

Andres Kaalep declares that he has no conflict of interest. Terez Sera has received travel grants and honoraria from EARL. Wim Oyen declares that he has no conflict of interest. Bernd J. Krause declares that he has no conflict of interest. Yan Liu declares that she has no conflict of interest. Arturo Chiti has received travel grants and speaker honoraria from General Electric Medical Systems. Ronald Boellaard is a member of the EARL scientific advisory board.

This article does not contain any studies with human participants or animals performed by any of the authors.

2.8 Supplementary material

IQQC	Mean recovery coefficient	Median	Standard Deviation	Skewness	% of submissions with RC below Earl specs	% of submissions with RC above Earl specs	% of submissions with RC within Earl specs
All 37mm sphere SUV _{max}	1.08 (± 0.01)	1.06	0.156	12.9	6%	13%	81%
Approved 37mm sphere SUV _{max}	1.05 (± 0.01)	1.04	0.160	18.5	2%	2%	96%
All 28mm sphere SUV _{max}	1.05 (± 0.01)	1.03	0.150	12.7	4%	13%	83%
Approved 28mm sphere SUV _{max}	1.03 (± 0.01)	1.02	0.154	18.9	0%	2%	98%
All 22mm sphere SUV _{max}	1.02 (± 0.01)	1.01	0.144	10.2	1%	17%	83%
Approved 22mm sphere SUV _{max}	1.00 (± 0.01)	0.99	0.140	17.9	0%	3%	97%
All 17mm sphere SUV _{max}	0.95 (± 0.01)	0.93	0.152	6.9	2%	23%	75%
Approved 17mm sphere SUV _{max}	0.92 (± 0.01)	0.91	0.138	14.1	1%	8%	92%
All 13mm sphere SUV _{max}	0.73 (± 0.01)	0.71	0.168	3.6	11%	14%	75%
Approved 13mm sphere SUV _{max}	0.70 (± 0.01)	0.69	0.123	11.9	6%	2%	92%
All 10mm sphere SUV _{max}	0.47 (± 0.01)	0.45	0.139	3.2	6%	32%	62%
Approved 10mm sphere SUV _{max}	0.45 (± 0.01)	0.44	0.108	8.8	3%	23%	74%

Supplemental table 1. IQQC SUV_{max} statistics for all sphere sizes calculated for all received and EARL approved submissions.

IQQC	Mean recovery coefficient	Median	Standard Deviation	Skewness	% of submissions below Earl specs	% of submissions above Earl specs	% of submissions within Earl specs
All 37mm sphere SUVmean	0.85 (± 0.01)	0.84	0.112	15.0	7%	18%	76%
Approved 37mm sphere SUVmean	0.83 (± 0.01)	0.83	0.120	19.4	2%	4%	94%
All 28mm sphere SUVmean	0.80 (± 0.01)	0.79	0.111	13.9	7%	17%	76%
Approved 28mm sphere SUVmean	0.79 (± 0.01)	0.79	0.116	19.6	2%	3%	95%
All 22mm sphere SUVmean	0.76 (± 0.01)	0.75	0.107	11.3	1%	32%	67%
Approved 22mm sphere SUVmean	0.75 (± 0.01)	0.74	0.106	18.7	0%	14%	86%
All 17mm sphere SUVmean	0.69 (± 0.01)	0.68	0.111	7.0	4%	25%	71%
Approved 17mm sphere SUVmean	0.67 (± 0.01)	0.67	0.101	14.4	1%	9%	89%
All 13mm sphere SUVmean	0.54 (± 0.01)	0.52	0.118	4.0	10%	17%	73%
Approved 13mm sphere SUVmean	0.52 (± 0.01)	0.52	0.088	12.9	5%	4%	90%
All 10mm sphere SUVmean	0.36 (± 0.01)	0.35	0.105	2.5	10%	30%	60%
Approved 10mm sphere SUVmean	0.34 (± 0.01)	0.34	0.082	7.2	7%	21%	73%

Supplemental table 2. IQQC SUVmean statistics for all sphere sizes calculated for all received and EARL approved submissions.

Supplemental Fig. 4. IQQC SUVmean (a to f) and SUVmax (g to l) recovery results, regular ongoing and all EARL approved submissions (pooled and per vendor). Dots represent EARL specification limits. Central line of the box is the median, edges of the box are the 25th and 75th percentiles, the whiskers extend to either the most extreme data points which are not considered outliers or 1.5 times interquartile range. The outliers are marked using plus signs.

Supplemental Fig. 5. Longitudinal analysis of IQQC results from 16 scanners. Recovery coefficients biases from the mean of the respective scanner. „Initial” data series represents the first IQQC submission for each system. Central line of the box is the median, edges of the box are the 25th and 75th percentiles, the whiskers extend to either the most extreme data points which are not considered outliers or 1.5 times interquartile range. The outliers are marked using plus signs.

Supplemental Fig. 6. Longitudinal analysis of the 16 systems' CalQC results. SUV bias values for each system are presented as separate lines. Dotted lines represent outliers and dashed lines subsequent corrective actions. Red dots represent datapoints outside EARL specifications.

Bibliography

1. Fletcher JW, Djulbegovic B, Soares HP, Siegel B a, Lowe VJ, Lyman GH, et al. Recommendations on the use of 18F-FDG PET in oncology. *J. Nucl. Med.* 2008;49:480–508.
2. de Geus-Oei L-F, van der Heijden HFM, Corstens FHM, Oyen WJG. Predictive and prognostic value of FDG-PET in nonsmall-cell lung cancer: a systematic review. *Cancer.* 2007;110:1654–64.
3. Shankar LK, Hoffman JM, Bacharach S, Graham MM, Karp J, Lammertsma AA, et al. Consensus Recommendations for the Use of 18F-FDG PET as an Indicator of Therapeutic Response in Patients in National Cancer Institute Trials. *J Nucl Med.* 2006;47:1059–66.
4. Gupta T, Master Z, Kannan S, Agarwal JP, Ghosh-Laskar S, Rangarajan V, et al. Diagnostic performance of post-treatment FDG PET or FDG PET/CT imaging in head and neck cancer: a systematic review and meta-analysis. *Eur. J. Nucl. Med. Mol. Imaging.* 2011;38:2083.
5. Ung YC, Maziak DE, Vanderveen JA, Smith CA, Gulenchyn K, Lacchetti C, et al. 18Fluorodeoxyglucose positron emission tomography in the diagnosis and staging of lung cancer: A systematic review. *J. Natl. Cancer Inst.* 2007;99:1753–67.
6. Lv Y-L, Yuan D-M, Wang K, Miao X-H, Qian Q, Wei S-Z, et al. Diagnostic performance of integrated positron emission tomography/ computed tomography for mediastinal lymph node staging in non-small cell lung cancer: a bivariate systematic review and meta-analysis. *J. Thorac. Oncol. International Association for the Study of Lung Cancer;* 2011;6:1350–8.
7. Eschmann SM, Friedel G, Paulsen F, Reimold M, Hehr T, Budach W, et al. 18F-FDG PET for assessment of therapy response and preoperative re-evaluation after neoadjuvant radio-chemotherapy in stage III non-small cell lung cancer. *Eur. J. Nucl. Med. Mol. Imaging.* 2007;34:463–71.
8. Hicks RJ, Kalf V, Macmanus MP, Ware RE, McKenzie AF, Matthews JP, et al. The Utility of F-FDG PET for Suspected Recurrent Non – Small Cell Lung Cancer After Potentially Curative Therapy : Impact on Management and Prognostic Stratification. 2001;42:10–21.
9. Dijkman BG, Schuurbiens OCJ, Vriens D, Looijen-Salamon M, Bussink J, Timmer-Bonte JNH, et al. The role of (18)F-FDG PET in the differentiation between lung metastases and synchronous second primary lung tumours. *Eur. J. Nucl. Med. Mol. Imaging.* Berlin/Heidelberg: Springer-Verlag; 2010;37:2037–47.
10. Hellwig D, Graeter TP, Ukena D, Groeschel A, Sybrecht GW, Schaefer HJ, et al. 18F-FDG PET for mediastinal staging of lung cancer: which SUV threshold makes sense? *J Nucl Med.* 2007;48:1761–6.
11. Agarwal M, Brahmanday G, Bajaj SK, Ravikrishnan KP, Wong C-YO. Revisiting the prognostic value of preoperative (18)F-fluoro-2-deoxyglucose ((18)F-FDG) positron emission tomography (PET) in early-stage (I & II) non-small cell lung cancers (NSCLC). *Eur. J. Nucl. Med. Mol. Imaging.* 2010;37:691–8.
12. Liao S, Penney BC, Wroblewski K, Zhang H, Simon CA, Kampalath R, et al. Prognostic value of metabolic tumor burden on 18F-FDG PET in nonsurgical patients with non-small cell lung cancer. *Eur. J. Nucl. Med. Mol. Imaging.* 2012;39:27–38.
13. Weber WA. Use of PET for Monitoring Cancer Therapy and for Predicting Outcome. *J. Nucl. Med.* 2005;46:983–95.
14. Hicks RJ. Role of 18F-FDG PET in Assessment of Response in Non-Small Cell Lung Cancer. *J. Nucl. Med.* 2009;50:31S–42.
15. Czernin J, Weber WA, Herschman HR. Molecular imaging in the development of cancer therapeutics. *Annu. Rev. Med. United States;* 2006;57:99–118.
16. Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. *Nat. Rev. Drug Discov. England;* 2003;2:566–80.
17. Weber WA. Assessing Tumor Response to Therapy. *J Nucl Med.* 2009;50:1S–10.
18. Bengtsson T, Hicks RJ, Peterson A, Port RE. 18F-FDG PET as a surrogate biomarker in non-small cell lung cancer treated with erlotinib: newly identified lesions are more informative than standardized uptake value. *J. Nucl. Med.* 2012;53:530–7.
19. Huang W, Zhou T, Ma L, Sun H, Gong H, Wang J, et al. Standard uptake value and metabolic tumor volume of 18F-FDG PET/CT predict short-term outcome early in the course of chemoradiotherapy in advanced non-small cell lung cancer. *Eur. J. Nucl. Med. Mol. Imaging.* 2011;38:1628.
20. Thie JA. Understanding the standardized uptake value, its methods, and implications for usage. *J.*

- Nucl. Med. 2004;45:1431–4.
21. Weber WA, Petersen V, Schmidt B, Tyndale-Hines L, Link T, Peschel C, et al. Positron emission tomography in non-small-cell lung cancer: prediction of response to chemotherapy by quantitative assessment of glucose use. *J. Clin. Oncol. United States*; 2003;21:2651–7.
 22. Kelly MD, Declerck JM. SUVref: reducing reconstruction-dependent variation in PET SUV. *EJNMMI Res.* 2011;1:16.
 23. Adams MC, Turkington TG, Wilson JM, Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. *AJR. Am. J. Roentgenol. United States*; 2010;195:310–20.
 24. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J. Nucl. Med.* 2009;50 Suppl 1:11S–20S.
 25. Kinahan PE, Fletcher JW. Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. *Semin. Ultrasound. CT. MR. United States*; 2010;31:496–505.
 26. Fahey FH, Kinahan PE, Doot RK, Kocak M, Thurston H, Poussaint TY. Variability in PET quantitation within a multicenter consortium. *Med. Phys. United States*; 2010;37:3660–6.
 27. Beyer T, Czernin J, Freudenberg LS. Variations in clinical PET/CT operations: results of an international survey of active PET/CT users. *J. Nucl. Med. United States*; 2011;52:303–10.
 28. Graham MM, Badawi RD, Wahl RL. Variations in PET/CT methodology for oncologic imaging at U.S. academic medical centers: an imaging response assessment team survey. *J. Nucl. Med.* 2011;52:311–7.
 29. Hristova I, Boellaard R, Vogel W, Mottaghy F, Marreard S, Collette S, et al. Retrospective quality control review of FDG scans in the imaging sub-study of PALETTE EORTC 62072/VEG110727: a randomized, double-blind, placebo-controlled phase III trial. *Eur. J. Nucl. Med. Mol. Imaging.* 2015;42:848–57.
 30. Boellaard R. Mutatis mutandis: harmonize the standard! *J. Nucl. Med.* 2012;53:1–3.
 31. Buckler AJ, Boellaard R. Standardization of quantitative imaging: the time is right, and 18F-FDG PET/CT is a good place to start. *J. Nucl. Med.* 2011;52:171–2.
 32. Sunderland JJ, Christian PE. Quantitative PET/CT Scanner Performance Characterization Based Upon the Society of Nuclear Medicine and Molecular Imaging Clinical Trials Network Oncology Clinical Simulator Phantom. *J Nucl Med.* 2015;56:145–52.
 33. Rausch I, Cal-González J, Dapra D, Gallowitsch HJ, Lind P, Beyer T, et al. Performance evaluation of the Biograph mCT Flow PET/CT system according to the NEMA NU2-2012 standard. *EJNMMI Phys.* 2015;2:26.
 34. Lasnon C, Desmots C, Quak E, Gervais R, Do P, Dubos-Arvis C, et al. Harmonizing SUVs in multicentre trials when using different generation PET systems: Prospective validation in non-small cell lung cancer patients. *Eur. J. Nucl. Med. Mol. Imaging.* 2013;40:985–96.
 35. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur. J. Nucl. Med. Mol. Imaging.* 2014;42:328–54.
 36. Delbeke D, Coleman RE, Guiberteau MJ, Brown ML, Royal HD, Siegel B a, et al. Procedure guideline for tumor imaging with 18F-FDG PET/CT 1.0. *J. Nucl. Med.* 2006;47:885–95.
 37. Graham MM, Wahl RL, Hoffman JM, Yap JT, Sunderland JJ, Boellaard R, et al. Summary of the UPICT protocol for 18F-FDG PET/CT imaging in oncology clinical trials. *J. Nucl. Med.* 2015;56:955–61.
 38. Makris NE, Huisman MC, Kinahan PE, AA, Boellaard R. Evaluation of strategies towards harmonization of FDG PET/CT studies in multicentre trials: Comparison of scanner validation phantoms and data analysis procedures. *Eur. J. Nucl. Med. Mol. Imaging.* 2013;40:1507–15.
 39. Scheuermann JS, Saffer JR, Karp JS, Levering AM, Siegel A. Qualification of PET Scanners for Use in Multicenter Cancer Clinical Trials: The American College of Radiology Imaging Network Experience. *J. Nucl. Med.* 2010;50:1187–93.
 40. Kinahan P, Wahl R, Shao L, Frank R, Perlman E. The QIBA profile for quantitative FDG-PET/CT oncology imaging. *J. Nucl. Med.* 2014;55:1520–.
 41. Boellaard R, Willemsen a T, Arends B, Visser EP. EARL procedure for assessing PET/CT system specific patient FDG activity preparations for quantitative FDG PET/CT studies. 2013. p. 1–3.
 42. EANM EARL website [Internet]. Available from: <http://earl.eanm.org>
 43. Geworski L, Knoop BO, de Wit M, Ivancevic V, Bares R, Munz DL. Multicenter Comparison of Calibration and Cross Calibration of PET Scanners. *J. Nucl. Med.* 2002;43:635–9.

Chapter 3

Feasibility of state of the art PET/CT systems performance harmonisation

Author(s): Andres Kaalep, Terez Sera, Sjoerd Rijnsdorp, Maqsood Yaqub,
Anne Talsma, Martin A. Lodge, Ronald Boellard

*as published in the European Journal of Nuclear Medicine and Molecular Imaging
(Kaalep et al, 2018)*

3.1 Abstract

Purpose: The objective of this study was to explore the feasibility of harmonising performance for PET/CT systems equipped with time-of-flight (ToF) and resolution modelling/point spread function (PSF) technologies. The second aim was producing a working prototype of new harmonising criteria with higher contrast recoveries than current EARL standards using various SUV metrics.

Methods: Four PET/CT systems with both ToF and PSF capabilities from 3 major vendors were used to acquire and reconstruct images of the NEMA NU2-2007 body phantom filled conforming EANM EARL guidelines. A total of 15 reconstruction parameter sets of varying pixel size, post filtering and reconstruction type, with 3 different acquisition durations were used to compare the quantitative performance of the systems. A target range for recovery curves was established such that it would accommodate the highest matching recoveries from all investigated systems. These updated criteria were validated on 18 additional scanners from 16 sites in order to demonstrate the scanners' ability to meet the new target range.

Results: Each of the four systems was found to be capable of producing harmonising reconstructions with similar recovery curves. The 5 reconstruction parameter sets producing harmonising results significantly increased SUVmean (25%) and SUVmax (26%) contrast recoveries compared with current EARL specifications. Additional prospective validation performed on 18 scanners from 16 EARL accredited sites demonstrated the feasibility of updated harmonising specifications. SUVpeak was found to significantly reduce the variability in quantitative results while producing lower recoveries in smaller (≤ 17 mm diameter) sphere sizes.

Conclusions: Harmonising PET/CT systems with ToF and PSF technologies from different vendors was found to be feasible. The harmonisation of such systems would require an update to the current multicentre accreditation program EARL in order to accommodate higher recoveries. SUVpeak should be further investigated as a noise resistant alternative quantitative metric to SUVmax.

3.2 Introduction

^{18}F -fluorodeoxyglucose (^{18}F -FDG) positron emission tomography (PET) and computed tomography (CT) hybrid imaging (PET/CT) is an important functional imaging tool being widely used for diagnosis, staging and therapy response evaluation in e.g. oncology [1–20]. Combined anatomical and functional information can be obtained in one session using hybrid PET/CT. In clinical practice, visual inspection of PET/CT images might be sufficient for the purposes of staging or restaging [7][21], however PET is a quantitative technique [22–26] and can provide more accurate and less observer-dependent metrics for diagnosis, therapy assessment and response monitoring using quantitative data in addition to visual interpretation [27]. In recent oncological clinical trials quantitative PET/CT data are also used for patient selection, stratification and therapy response monitoring. However, variability, reproducibility and accuracy of quantitative PET/CT imaging [28–34] have to be considered. Scientific societies such as the European Association of Nuclear Medicine (EANM), American College of Radiology (ACR), American Association of Physicists in Medicine (AAPM), Radiological Society of North America (RSNA) and Society of Nuclear Medicine and Molecular Imaging (SNMMI) are closely collaborating to promote standardisation of practices in order to reduce variability of quantification in multicentre clinical trials. Initiatives such QIBA-UPICT, SNMMI-CTN and EANM-EARL are providing quality control programs to assure quantitative comparability [35–40].

High utilization of PET/CT in oncology can be attributed to the availability of ^{18}F -FDG [5,41]. Dynamic PET scans and pharmacokinetic modelling to evaluate the rate of glucose metabolism of tumour is an excellent method for quantification [27] but the technical impediments such as the limited scanner field of view and increased scan acquisition time make it unfeasible for routine use [42]. In clinical practice, a simplified uptake metric such as the standard uptake value (SUV) [43,44] is therefore most commonly used. While SUV analysis is relatively easy to apply, it suffers from multiple technical, physical and biological factors that can significantly affect quantification [27]. The required level of harmonisation depends on the intended use of the PET study. When the same PET/CT system is used for therapy assessment and based on relative changes in SUV before and after therapy, a high reproducibility rather than absolute accuracy might be most important. It has been shown that in this case, when the scanner performance remains unchanged over time, consistent application of a certain methodology could be sufficient [34,45]. However, patients are often scanned on different PET/CT systems, either because the scanner had been replaced by a new one, or in different institutions, which makes accurate cross-calibration of systems a crucial requirement. Absolute quantitative measures (e.g., residual uptake of ^{18}F -FDG after therapy session) are also being used for differentiation between malignant and benign lesions, determining prognosis and response monitoring [27]. This again requires high reproducibility and comparability of the quantitative data, especially in multicentre settings.

One of the challenges in PET/CT systems performance harmonisation is the variability caused by different PET/CT technologies available in the field. Multicentre standards should not be based on the less performing systems; they need to fit with the highest, yet common denominator in systems' performances. Additionally, in case of optimization of PET/CT systems performance for lesion detection, a single centre quantification doesn't necessarily coincide with a multicentre one. A particular challenge for recent PET/CT systems resulted from the introduction of time-of-flight (ToF) and resolution modelling (point spread function (PSF)) capabilities. The latter increased tumour detectability but also caused higher variability across centres, since some have and others lack these technologies. Currently a large number of the EARL accredited PET/CT systems [46] do not have PSF image reconstruction capabilities. However, it is expected that over the next couple of years the majority of the PET/CT systems will be equipped with these new reconstruction techniques.

The aim of this paper is to explore the feasibility of harmonising performance of PET/CT systems equipped with the latest PET technologies such as TOF and PSF, which were recently commercially released.

3.3 Materials and methods

3.3.1 PET/CT system selection

Four PET/CT systems equipped with both ToF and PSF capabilities from 3 major vendors (General Electric (GE), Siemens and Philips) were selected for this study. Systems included were Siemens Biograph mCT (Siemens system 1), Siemens Biograph mCT Flow (Siemens system 2), GE Discovery 710 (GE system) and Philips Ingenuity TF 128 (Philips system). The equipment was calibrated in accordance with the corresponding manufacturer's instructions. In addition, all systems were participating and accredited in the EANM/EARL ^{18}F -FDG PET/CT accreditation program. Detailed specifications for the systems can be found in supplemental table 1 and references [47][48][49][50][51].

3.3.2 Phantom experiments

The phantoms and filling procedures used complied with the EANM/EARL guidelines for Image Quality QC measurements which need to be performed annually as part of the EANM/EARL accreditation program [35]. The NEMA NU2-2007 body phantom was used, which is

a plastic cylinder in form of a fillable torso cavity, to act as a background compartment. It has a 5 cm diameter cylindrical lung insert in the centre and 6 fillable spheres with internal diameters of 10 mm, 13 mm, 17 mm, 22 mm, 28 mm and 37 mm, positioned coaxially around the lung insert. The lung insert is filled with polystyrene beads in order to mimic lung tissue. The phantom background compartment and the spherical inserts were filled with ^{18}F -FDG solutions aimed at activity concentrations of 2 kBq/mL and 20 kBq/mL, respectively, at the start of the measurements, resulting in a sphere to background activity concentration ratio of 10:1.

3.3.3 Acquisition & reconstruction parameters

In accordance with current EANM/EARL guidelines for ^{18}F -FDG Image Quality QC phantom imaging [35], a low dose CT acquisition, followed by an emission scan consisting of two bed positions with an acquisition time of 5 minutes per bed position is to be acquired for the “image quality” dataset to assess contrast recovery performance. In this study, acquisition time of 5 minutes per bed position was selected as the reference for high count statistics. In order to investigate the effect of reduced count statistics on contrast recovery, data acquired with shorter acquisition times, respectively 2 and 1 minutes per bed position, were collected. The GE and Philips systems had list mode data acquisition capability available, which meant that only the 5 minute/bed position emission scans were acquired and reconstructions with shorter acquisition times were generated retrospectively from the list mode data. On the Siemens systems included in this study, multiple shorter emission scans were acquired with the phantom left in an unchanged position. In order to facilitate the Siemens Flow system’s (Siemens system 2) possibility of performing scanning with continuous table movement, instead of a specific bed position scanning duration, table feed speeds of 0.5 mm/s, 1 mm/s and 2 mm/s were selected, resulting in similar acquisition times as with the other scanners.

Reconstructions were performed using the software available on each of the PET/CT systems. TOF, PSF, normalisation, randoms, scatter and attenuation corrections were applied and the reconstruction parameters were selected to increase overall contrast recovery, meanwhile aiming at achieving comparable recovery values across systems (for each sphere). In addition, we also considered to achieve comparable recovery values between the spheres to minimize severe partial volume effects as well as large Gibbs overshoots. Clinically used and vendor recommended reconstruction parameters were applied and varied. 3 iterations with 21 subsets were used for Siemens 1 (Biograph mCT) and 2 iterations with 21 subsets for Siemens 2 (mCT Flow) reconstruction. For GE - B, D, F and G (Discovery 710) - 2 iterations with 24 subsets and VPFXS reconstruction method, while for GE - A, C and E - QCFX reconstruction method, with an unknown number of iterations and subsets, was used. For the Philips systems the iterations/subsets were 3/33 but these could not be selected prior scanning, neither its values retrieved from the DICOM header of the images; BLOB OS TF reconstruction method was used.

Reconstruction mode	Post filter width (mm)	Q.Clear β value	Pixel size (mm)	Slice thickness (mm)	Long frame duration (s)	Medium frame duration (s)	Short frame duration (s)
GE - A	N/A	200	2.73	3.27	300	120	60
GE - B	0	N/A	2.73	3.27	300	120	60
GE - C	N/A	350	2.73	3.27	300	120	60
GE - D	3	N/A	2.73	3.27	300	120	60
GE - E	N/A	800	2.73	3.27	300	120	60
GE - F	5	N/A	2.73	3.27	300	120	60
GE - G	6.4	N/A	2.73	3.27	300	120	60
Philips - A	N/A	N/A	2.00	2.00	301	120	60
Philips - B	N/A	N/A	4.00	4.00	301	120	60
Siemens 1 - A	0	N/A	2.04	2.00	300	120	60
Siemens 1 - B	0	N/A	1.59	2.00	300	120	60
Siemens 1 - C	3	N/A	2.04	2.00	300	120	60
Siemens 1 - D	5	N/A	2.04	2.00	300	120	60
Siemens 1 - E	6.5	N/A	3.18	2.00	300	120	60
Siemens 2 - A	5	N/A	4.07	5.00	223	111	56

Table 1. Acquisition and reconstruction settings for initial 15 reconstruction modes

Different Gaussian filters and pixel sizes within clinically relevant range were also investigated in order to study their effects on contrast recovery. Additionally, for the GE system, a proprietary reconstruction method, the “Q.Clear”, which uses a Bayesian penalized-likelihood reconstruction algorithm, was investigated using different penalization factors (β) and its effect on quantitative image quality was evaluated. Due to differences among vendors and models, the available reconstruction parameters and their ranges were limited based on availability and/or user selectability. In total, 15 reconstruction parameter sets (reconstruction modes) were used to assess and compare the quantitative performance of the investigated systems. Each reconstruction mode was applied on three different scans, acquired with long (~4 min/bed for Siemens Flow system; ~5 min/bed for all other systems), medium (~2 min/bed) and short (~1 min/bed) frame durations. A summary of the acquisition and reconstruction settings of the 15 reconstruction modes is presented in table 1.

3.3.4 Data analysis

Data reconstructed on the PET/CT were exported to a PC for further analysis using the EARL semi-automatic tool [35] designed for quantitative analysis of images of NEMA NU2-2007 body phantom, filled conforming EANM/EARL guidelines for ^{18}F -FDG Image Quality QC phantom imaging. The software tool requires phantom images in DICOM format and filling data as input, and extracts SUV recovery for the spheres, calibration factor for the background compartment and standard deviation and coefficient of variation from uniform images of the background. SUV recovery coefficient (RC) is defined as the ratio between measured and expected activity concentration in each spherical insert. RC values were calculated based on 50% background corrected isocontour VOI ($\text{RC}_{\text{SUVmean}}$), maximum voxel value included in VOI ($\text{RC}_{\text{SUVmax}}$) and spherical VOI with a diameter of 12 mm, positioned so to yield the highest uptake ($\text{RC}_{\text{SUVpeak}}$) [39][35][52].

Prior to further analysis, all data were corrected for system calibration bias in order to be able to compare the various reconstruction modes' impact on RCs and not to be effected by inter-scanner calibration errors. For this purpose, to all RCs a correction factor, defined as the ratio between expected and measured activity concentration in the corresponding uniform background compartment was applied. For the 15 initial reconstruction modes, inter-scanner global correction factors ranged from 0.88 to 1.12, with the mean and standard deviation being 0.98 and 0.055, respectively. Intra-scanner changes were below 1%. For the 23 additional reconstructions, the inter-scanner global correction factors ranged from 0.93 to 1.10 (one system, however, showed a correction factor of 0.8), with the mean and standard deviation values of 0.99 and 0.055, respectively.

3.3.5 Selection of harmonising reconstruction modes

The primary objective of this study was to find reconstruction modes providing high, yet uniform contrast recoveries within the spheres of the NEMA NU2-2007 body phantom, which could be matched across all generations of PET/CT systems currently used in clinical practice – which would result in quantitative harmonisation of PET/CT systems.

$\text{RC}_{\text{SUVmean}}$, $\text{RC}_{\text{SUVmax}}$ and $\text{RC}_{\text{SUVpeak}}$ curves for all reconstructed phantom images were plotted against sphere diameters (figure 1) and characterised using visual and quantitative analysis, for which the applied metrics are summarised in table 2. Reconstruction modes with higher RCs than current EARL specifications, as well as tightly grouped and stable $\text{RC}_{\text{SUVmean}}$ and $\text{RC}_{\text{SUVmax}}$ curves, were sought for harmonisation purposes.

Metric	Description of metric
SUV_{mean}	Ratio of image derived average radioactivity concentration within a region of interest and the whole body concentration of the injected radioactivity
SUV_{max}	Ratio of image derived maximum (single pixel) radioactivity concentration within a region of interest and the whole body concentration of the injected radioactivity
SUV_{peak}	Ratio of image derived average radioactivity concentration within a 12 mm diameter spherical volume within the region of interest, positioned to yield the highest uptake, and the whole body concentration of the injected radioactivity
RC	Recovery Coefficient - the ratio between image derived and expected activity concentration
MCR*	Mean Contrast Recovery - mean RC of all spheres in corresponding reconstruction mode's long duration acquisition. Parameter is indicative of reconstruction mode's overall contrast recovery potential.
CoVMCR	Coefficient of Variation (SD/mean*100, %) of a group of MCR values. Parameter is indicative of RC curves' alignment within a group.
CRV_{medium}^*	Contrast Recovery Variability - Mean deviation of medium duration acquisition spheres' RCs from the corresponding values of long duration acquisition.
CRV_{short}^*	Contrast Recovery Variability - Mean deviation of short duration acquisition spheres' RCs from the corresponding values of long duration acquisition.
CoV_{BG}^*	Coefficient of Variation (SD/mean *100, %) of measured activity concentration within the uniform background compartment of the phantom. Parameter is indicative of the noise present in the images.
Curvature	Long acquisition duration root-mean-square deviation of spheres' RC values from RC value of the largest (37 mm) sphere. Parameter characterises the deviation of smaller spheres' RC values which usually cause the RC-object size relation to assume a curved shape.
Absolute error	Long acquisition duration root-mean-square deviation of spheres' RC values from unity. The parameter characterises the reconstruction mode's ability to report accurate activity concentration values.
Curvature (excl. 10 mm sphere)	Same as „curvature“ but excluding the smallest (10 mm) sphere.
Absolute error (excl. 10 mm sphere)	Same as „absolute error“ but excluding the smallest (10 mm) sphere.

*Quantitative metrics that were retrospectively used to determine harmonising cut-off criteria

Table 2. Description of quantitative metrics used.

Fig. 1. RC curves derived from 15 initial reconstruction modes using SUVmean (a), SUVmax (b) and SUVpeak (c) quantitative metrics. Only long acquisition time frame curves are displayed. GE (Q.Clear) – blue dashed lines, GE (non-Q.Clear) – blue solid lines, Philips – red solid lines, Siemens 1 – orange solid lines, Siemens 2 – green solid lines, current EARL specifications – black solid lines.

The harmonising reconstruction modes were selected by simultaneously analysing quantitative characteristics of the reconstruction modes along with visual appearance of the RC curves. Following considerations were kept in mind while determining feasible reconstruction modes – (1) the proposed harmonising specifications should provide an increase over the current EARL compliant RC values, (2) the bandwidth of RCs should be similar to the current Earl specification limits and (3) the harmonising RC curves should not demonstrate major overshoots (=upward bias) due to Gibbs artefacts. While the harmonising reconstruction modes were selected based on the abovementioned considerations, quantitative cut-off criteria were retrospectively determined and stated in Table 9 based on the bandwidth and characteristics of harmonising reconstruction modes. Performances of the candidate reconstruction modes were compared with the initial group of reconstructions as well as current EARL accreditation specifications.

3.3.6 Mean Contrast Recovery (MCR)

Mean contrast recovery (MCR) was calculated in order to evaluate overall contrast recovery potential of a reconstruction mode while Coefficient of Variation of the MCR parameter (CoV_{MCR}) was used to characterise agreement among various reconstruction modes' RC curves. Increased coinciding MCR and reduced CoV_{MCR} values were preferred.

3.3.7 Contrast Recovery Variability (CRV)

Contrast Recovery Variability ($\text{CRV}_{\text{medium}}$ and $\text{CRV}_{\text{short}}$) parameters were used to evaluate a reconstruction mode's ability to produce consistent results in case of reduced count statistics. In order to achieve it, RCs of short and medium time frame acquisitions were compared to the long acquisition's corresponding spheres' RCs and relative differences calculated. Lower values were deemed preferable as being indicative of reconstruction mode's stability and reduced variability in noisy environments.

3.3.8 Noise

Image noise was quantitatively evaluated by measuring the Coefficient of Variation ($\%, \text{SD}/\text{Mean} \times 100$) in the uniform background compartment (CoV_{BG}) for each reconstruction mode and acquisition time frame. CoV_{BG} cut-off limit of 15%, based on the existing EARL guideline and UPICT [35,37,40], was implemented to determine suitable reconstruction modes for harmonisation. Reconstruction modes providing lower noise images were deemed preferable.

Fig. 2. RC curves derived from suggested harmonising reconstruction modes using SUVmean (a), SUVmax (b) and SUVpeak (c) quantitative metrics along with current EARL and possible new specifications. Only long acquisition time frame curves are displayed. GE (Q.Clear) – blue dashed lines, GE (non-Q.Clear) – blue solid lines, Philips – red solid lines, Siemens 1 – orange solid lines, Siemens 2 – green solid lines, current EARL specifications – black solid lines, possible new EARL specifications – black dashed lines.

3.3.9 Curvature and absolute error

Curvature and absolute error parameters were used to evaluate RC variability and absolute accuracy of RC measurements due to changes in sphere/lesion size. Reduced values were preferable but similar magnitude across systems/reconstructions was given priority.

3.3.10 Visual analysis

Visual analysis of the RC curves was used to identify reconstruction modes that exhibited abnormal behaviour or localised variations, such as exaggerated Gibbs artefacts, that were not identified by the previously described quantitative parameters. The reconstruction modes, which were considered for harmonisation based on SUVmean and SUVmax performance, were also used to develop provisional specifications for SUVpeak.

3.3.11 Validation of reconstruction modes for harmonisation

In order to prospectively evaluate the reproducibility and inter-scanner variability of the proposed reconstruction modes for harmonisation, 16 EARL accredited facilities, equipped with current generation PET/CT systems, participated in the study and provided the requested reconstructions from independent phantom acquisitions applying acquisition and reconstruction parameters (supplemental table 2) identical or similar to the reconstructions proposed for harmonisation purposes. Data received from the centres was analysed in the same way as the reconstructions in the pilot study.

3.4 Results

3.4.1 New specifications proposed for harmonisation

Analysis of the initial 15 reconstruction modes resulted in 5 reconstruction modes, which produced highest uniform contrast recoveries and were feasible for all of the investigated systems considering SUVmean and SUVmax (Philips - B, GE - E, GE - F, Siemens 1 - D and Siemens 2 - A), to be considered for harmonisation. In order to accommodate unavoidable inter-scanner variability and reproducibility errors due to equipment calibration and user inaccuracy, all of the RC ranges were expanded to be proportional (i.e. using the same bandwidth of performance,

but taking into account increased contrast recovery) to current EARL specifications for sphere recoveries. Specifications, based on the current findings, proposed for harmonisation along with current EARL specifications are presented in Table 13. Bandwidths for proposed and current EARL specifications as well as the RC curves derived from the 5 reconstruction modes are presented in figure 2. For the provisional SUVpeak specifications, average sphere recoveries of the 5 reconstruction modes and a bandwidth of ± 2 standard deviations was used. Additionally, recovery coefficients are plotted as function of background noise for each sphere and per SUVmetric (presented in supplemental figures 4-6). Axial slices of the phantom data from the 5 harmonising reconstructions are shown in supplemental figure 7.

3.4.2 Mean Contrast Recovery (MCR)

SUVmean and SUVmax RC curves vary substantially among different systems and reconstruction modes as seen in Figure 1 and Tables 3 and 4. The reconstruction mode showing the lowest recoveries (Siemens 1 – E) produced a SUVmean MCR value of 0.714 and SUVmax MCR of 0.948 while for the highest recovery reconstruction mode (Siemens 1 – A), the corresponding values were 1.09 and 1.56 – a difference of more than 50%. SUVpeak MCR values were found to be between 0.754 and 0.929. CoVMCR values for the 15 reconstruction modes were 12.4% and 15.4% for SUVmean and SUVmax, respectively, while for SUVpeak, CoVMCR was 6.0%.

For the 5 reconstruction modes proposed for harmonisation, the range of MCR values were 0.770 – 0.816 and 1.01 – 1.09 for SUVmean and SUVmax, respectively. The harmonising reconstruction modes produced SUVpeak MCR values in the range of 0.784 – 0.823. CoVMCR values for SUVmean, SUVmax and SUVpeak were 2.2%, 2.9% and 2.2%, respectively.

3.4.3 Contrast Recovery Variability (CRV)

The initial 15 reconstruction modes demonstrated variable sensitivity in function of count statistics. The expected increase in variability with decrease in count statistics was observed in all reconstruction modes by comparing CRVmedium and CRVshort values (Tables 3-5). The CRVmedium results for SUVmean, SUVmax and SUVpeak ranged from 2.4% to 8.4%, 2.7% to 17.8% and 1.6% to 4.5%, respectively. The CRVshort results for SUVmean, SUVmax and SUVpeak ranged from 2.3% to 14.5%, 4.9% to 20.4% and 2.7% to 6.3%, respectively.

For the 5 reconstruction modes proposed for harmonisation, the CRVmedium results for SUVmean, SUVmax and SUVpeak ranged from 2.7% to 5.3%, 3.7% to 8.0% and 2.8% to 3.0%, respectively. The CRVshort results for SUVmean, SUVmax and SUVpeak ranged from 2.3% to 6.2%, 5.2% to 9.2% and 2.9% to 5.8%, respectively.

3.4.4 Noise

The CoVBG values are summarised in supplemental figure 8. The average CoVBG of all reconstruction modes with long time frame was 12.6%. For medium and short acquisition times, the corresponding values were 19.7% and 27.0%, respectively. The selected reconstruction modes for harmonisation purposes produced average CoVBG values of 9.4%, 14.0% and 18.4% for long, medium and short acquisition time frames, respectively.

3.4.5 Curvature and absolute error

Curvatures for the initial 15 reconstruction modes were in the ranges of 0.031 – 0.269, 0.076 – 0.290 and 0.305 – 0.413 for SUVmean, SUVmax and SUVpeak, respectively. For the 5 reconstruction modes suggested for harmonisation the SUVmean, SUVmax and SUVpeak curvatures were in the ranges of 0.197 – 0.236, 0.212 – 0.267 and 0.350 – 0.404, respectively.

Absolute errors for the initial 15 reconstruction modes were in the ranges of 0.053 – 0.329, 0.157 – 0.566 and 0.232 – 0.367 for SUVmean, SUVmax and SUVpeak, respectively. For the 5 reconstruction modes selected for harmonisation, the SUVmean, SUVmax and SUVpeak curvatures ranged between 0.222 – 0.277, 0.170 – 0.232 and 0.291 – 0.346, respectively.

3.4.6 Visual analysis

Significant variations in investigated RC curves' shapes and positions of Siemens 1 - A, B, C, GE - A, B, C, D and Philips – A reconstruction modes were noticed when compared with other systems or acquisition times and considered unsuitable for harmonisation.

Reconstruction mode	MCR	CRV _{medium}	CRV _{short}	Curvature	Absolute error	Curvature (excl. 10 mm sphere)	Absolute error (excl. 10 mm sphere)
GE - A	0.956	5.0%	6.8%	0.031	0.053	0.023	0.040
GE - B	0.903	8.4%	8.7%	0.139	0.147	0.022	0.050
GE - C	0.887	6.7%	6.4%	0.109	0.140	0.025	0.077
GE - D	0.859	6.3%	6.6%	0.168	0.188	0.053	0.092
GE - E	0.806	5.2%	6.2%	0.218	0.253	0.075	0.134
GE - F	0.770	5.3%	5.3%	0.228	0.277	0.120	0.183
GE - G	0.725	3.8%	4.9%	0.253	0.321	0.147	0.228
Philips - A	0.845	3.3%	4.2%	0.149	0.192	0.088	0.134
Philips - B	0.800	2.7%	2.3%	0.236	0.271	0.124	0.165
Siemens 1 - A	1.086	6.0%	14.5%	0.097	0.117	0.108	0.125
Siemens 1 - B	1.038	3.8%	12.5%	0.072	0.076	0.071	0.081
Siemens 1 - C	0.952	3.3%	8.3%	0.111	0.101	0.048	0.043
Siemens 1 - D	0.816	2.9%	5.1%	0.197	0.222	0.097	0.138
Siemens 1 - E	0.714	2.4%	4.0%	0.269	0.329	0.166	0.238
Siemens 2 - A	0.804	3.0%	4.4%	0.203	0.238	0.100	0.150
Min	0.714	2.4%	2.3%	0.031	0.053	0.022	0.040
Max	1.086	8.4%	14.5%	0.269	0.329	0.166	0.238
Average	0.864	4.5%	6.7%	0.165	0.195	0.085	0.125
COV _{MCR}	12.4%						

Table 3. Analysis results of 15 initial reconstruction modes using SUV_{mean} quantitative metric. Values found to be outside of acceptable range during retrospective quantitative analysis, are coloured red.

Reconstruction mode	MCR	CRV _{medium}	CRV _{short}	Curvature	Absolute error	Curvature (excl. 10 mm sphere)	Absolute error (excl. 10 mm sphere)
GE - A	1.245	17.8%	20.4%	0.081	0.255	0.089	0.265
GE - B	1.201	11.9%	19.7%	0.160	0.236	0.052	0.257
GE - C	1.142	12.9%	15.1%	0.076	0.157	0.036	0.172
GE - D	1.139	10.6%	15.8%	0.181	0.194	0.047	0.200
GE - E	1.036	7.2%	7.7%	0.212	0.178	0.041	0.119
GE - F	1.013	8.0%	9.2%	0.235	0.170	0.085	0.099
GE - G	0.951	5.5%	6.6%	0.274	0.203	0.129	0.094
Philips - A	1.146	7.2%	15.0%	0.176	0.204	0.103	0.218
Philips - B	1.061	3.7%	5.2%	0.267	0.232	0.150	0.197
Siemens 1 - A	1.555	10.1%	20.3%	0.126	0.566	0.139	0.574
Siemens 1 - B	1.477	8.0%	19.1%	0.116	0.487	0.112	0.505
Siemens 1 - C	1.325	5.4%	12.5%	0.148	0.346	0.104	0.375
Siemens 1 - D	1.094	3.9%	7.9%	0.218	0.179	0.080	0.165
Siemens 1 - E	0.948	2.7%	4.9%	0.290	0.199	0.145	0.084
Siemens 2 - A	1.045	3.7%	5.4%	0.246	0.184	0.104	0.138
Min	0.948	2.7%	4.9%	0.076	0.157	0.036	0.084
Max	1.555	17.8%	20.4%	0.290	0.566	0.150	0.574
Average	1.159	7.9%	12.3%	0.187	0.253	0.094	0.231
COV _{MCR}	15.4%						

Table 4. Analysis results of 15 initial reconstruction modes using SUV_{max} quantitative metric. Values found to be outside of acceptable range during retrospective quantitative analysis, are coloured red.

Reconstruction mode	MCR	CRV _{medium}	CRV _{short}	Curvature	Absolute error	Curvature (excl. 10 mm sphere)	Absolute error (excl. 10 mm sphere)
GE - A	0.848	3.9%	3.7%	0.334	0.287	0.187	0.153
GE - B	0.833	3.4%	5.7%	0.381	0.310	0.237	0.179
GE - C	0.840	2.3%	3.6%	0.359	0.302	0.211	0.166
GE - D	0.823	3.9%	6.3%	0.389	0.320	0.248	0.191
GE - E	0.821	2.9%	4.1%	0.400	0.339	0.250	0.203
GE - F	0.784	3.3%	5.8%	0.404	0.346	0.272	0.223
GE - G	0.757	3.1%	5.9%	0.413	0.367	0.287	0.248
Philips - A	0.874	3.2%	3.4%	0.328	0.281	0.192	0.161
Philips - B	0.796	2.8%	2.9%	0.383	0.341	0.263	0.229
Siemens 1 - A	0.901	4.5%	6.3%	0.305	0.232	0.148	0.090
Siemens 1 - B	0.929	1.6%	4.2%	0.325	0.240	0.154	0.103
Siemens 1 - C	0.872	3.3%	5.0%	0.308	0.251	0.151	0.107
Siemens 1 - D	0.823	3.0%	4.5%	0.350	0.291	0.204	0.155
Siemens 1 - E	0.754	3.9%	2.7%	0.382	0.346	0.255	0.226
Siemens 2 - A	0.789	2.9%	4.9%	0.355	0.323	0.240	0.214
Min	0.754	1.6%	2.7%	0.305	0.232	0.148	0.090
Max	0.929	4.5%	6.3%	0.413	0.367	0.287	0.248
Average	0.830	3.2%	4.6%	0.361	0.305	0.220	0.177
COV _{MCR}	6.0%						

Table 5. Analysis results of 15 initial reconstruction modes using SUV_{peak} quantitative metric.

Reconstruction mode	MCR	CRV _{medium}	CRV _{short}	Curvature	Absolute error	Curvature (excl. 10 mm sphere)	Absolute error (excl. 10 mm sphere)
GE - E	0.806	5.2%	6.2%	0.218	0.253	0.075	0.134
GE - F	0.770	5.3%	5.3%	0.228	0.277	0.120	0.183
Philips - B	0.800	2.7%	2.3%	0.236	0.271	0.124	0.165
Siemens 1 - D	0.816	2.9%	5.1%	0.197	0.222	0.097	0.138
Siemens 2 - A	0.804	3.0%	4.4%	0.203	0.238	0.100	0.150
Min	0.770	2.7%	2.3%	0.197	0.222	0.075	0.134
Max	0.816	5.3%	6.2%	0.236	0.277	0.124	0.183
Average	0.799	3.8%	4.6%	0.216	0.252	0.103	0.154
COV_{MCR}	2.2%						
EARL min	0.570	N/A	N/A	0.282	0.466	0.198	0.393
EARL max	0.710	N/A	N/A	0.277	0.342	0.176	0.251
Earl Average	0.640	N/A	N/A	0.279	0.403	0.187	0.321

Table 6. Results of the analysis of 5 reconstruction modes considered for harmonisation using SUV_{mean} quantitative metric.

Reconstruction mode	MCR	CRV _{medium}	CRV _{short}	Curvature	Absolute error	Curvature (excl. 10 mm sphere)	Absolute error (excl. 10 mm sphere)
GE - E	1.036	7.2%	7.7%	0.212	0.178	0.041	0.119
GE - F	1.013	8.0%	9.2%	0.235	0.170	0.085	0.099
Philips - B	1.061	3.7%	5.2%	0.267	0.232	0.150	0.197
Siemens 1 - D	1.094	3.9%	7.9%	0.218	0.179	0.080	0.165
Siemens 2 - A	1.045	3.7%	5.4%	0.246	0.184	0.104	0.138
Min	1.013	3.7%	5.2%	0.212	0.170	0.041	0.099
Max	1.094	8.0%	9.2%	0.267	0.232	0.150	0.197
Average	1.050	5.3%	7.1%	0.236	0.189	0.092	0.144
COV_{MCR}	2.9%						
EARL min	0.730	N/A	N/A	0.347	0.355	0.220	0.237
EARL max	0.970	N/A	N/A	0.339	0.236	0.176	0.121
Earl Average	0.850	N/A	N/A	0.342	0.277	0.198	0.142

Table 7. Results of the analysis of 5 reconstruction modes considered for harmonisation using SUV_{max} quantitative metric.

Reconstruction mode	MCR	CRV _{medium}	CRV _{short}	Curvature	Absolute error	Curvature (excl. 10 mm sphere)	Absolute error (excl. 10 mm sphere)
GE - E	0.821	2.9%	4.1%	0.400	0.339	0.250	0.203
GE - F	0.784	3.3%	5.8%	0.404	0.346	0.272	0.223
Philips - B	0.796	2.8%	2.9%	0.383	0.341	0.263	0.229
Siemens 1 - D	0.823	3.0%	4.5%	0.350	0.291	0.204	0.155
Siemens 2 - A	0.789	2.9%	4.9%	0.355	0.323	0.240	0.214
Min	0.784	2.8%	2.9%	0.350	0.291	0.204	0.155
Max	0.823	3.3%	5.8%	0.404	0.346	0.272	0.229
Average	0.803	3.0%	4.4%	0.378	0.328	0.246	0.205
COV _{MCR}	2.2%						

Table 8. Results of the analysis of 5 reconstruction modes considered for harmonisation using SUV_{peak} quantitative metric.

	SUV _{mean}	SUV _{max}
MCR	±11% (0.77-0.96)	±13% (1.01-1.31)
CRV _{medium}	6%	8%
CRV _{short}	7%	9%
Visual analysis	No excessive Gibbs and partial volume artefacts	
Noise	Background CoV ≤15% (high statistics acquisition)	

Table 9. Retrospectively determined quantitative cut-off criteria for the harmonising reconstructions.

Reconstruction mode	MCR	CRV _{medium}	CRV _{short}	Curvature	Absolute error	Curvature (excl. 10 mm sphere)	Absolute error (excl. 10 mm sphere)
Ingenuity 1	0.820	N/A	N/A	0.213	0.249	0.106	0.145
Ingenuity 2	0.694	N/A	N/A	0.276	0.365	0.164	0.263
mCT Flow 1	0.691	N/A	N/A	0.303	0.368	0.196	0.270
mCT Flow 2	0.711	N/A	N/A	0.298	0.339	0.190	0.242
mCT Flow 3	0.816	N/A	N/A	0.193	0.231	0.079	0.136
mCT 1	0.847	N/A	N/A	0.176	0.194	0.080	0.112
mCT 2	0.786	N/A	N/A	0.194	0.250	0.115	0.181
mCT 3	0.825	N/A	N/A	0.188	0.208	0.113	0.142
mCT 4	0.765	N/A	N/A	0.174	0.262	0.091	0.195
mCT 5	0.786	N/A	N/A	0.195	0.245	0.119	0.179
mCT 6	0.811	N/A	N/A	0.136	0.207	0.078	0.161
Discovery 710 1	0.847	N/A	N/A	0.153	0.182	0.079	0.120
Discovery 710 2	0.793	N/A	N/A	0.217	0.254	0.129	0.174
Discovery 710 1 Q.Clear 1	0.887	N/A	N/A	0.120	0.145	0.027	0.074
Discovery 710 2 Q.Clear 2	0.817	N/A	N/A	0.211	0.236	0.110	0.146
Discovery 710 3 Q.Clear 3	0.895	N/A	N/A	0.121	0.144	0.042	0.073
GE Discovery MI 1	0.794	N/A	N/A	0.150	0.228	0.099	0.182
GE Discovery MI 2	0.813	N/A	N/A	0.171	0.214	0.102	0.155
GE Discovery MI 1 Q.Clear 1	0.857	N/A	N/A	0.081	0.151	0.055	0.129
GE Discovery MI 2 Q.Clear 2	0.869	N/A	N/A	0.118	0.156	0.039	0.096
GE Discovery IQ 1	0.817	N/A	N/A	0.219	0.244	0.077	0.123
GE Discovery IQ 1 Q.Clear 1	0.818	N/A	N/A	0.221	0.246	0.069	0.118
Vereos 1	0.757	N/A	N/A	0.191	0.277	0.087	0.195
Min	0.691			0.081	0.144	0.027	0.073
Max	0.895			0.303	0.368	0.196	0.270
Average	0.805			0.188	0.235	0.098	0.157
COV _{MCR}	6.6%						

Table 10. Analysis results of 23 additional reconstructions using SUVmean quantitative metric.

Reconstruction mode	MCR	CRV _{medium}	CRV _{short}	Curvature	Absolute error	Curvature (excl. 10 mm sphere)	Absolute error (excl. 10 mm sphere)
Ingenuity 1	1.094	N/A	N/A	0.278	0.264	0.143	0.228
Ingenuity 2	0.917	N/A	N/A	0.334	0.288	0.188	0.167
mCT Flow 1	0.911	N/A	N/A	0.347	0.270	0.207	0.159
mCT Flow 2	0.943	N/A	N/A	0.350	0.234	0.187	0.109
mCT Flow 3	1.071	N/A	N/A	0.237	0.211	0.110	0.179
mCT 1	1.118	N/A	N/A	0.185	0.179	0.057	0.179
mCT 2	1.038	N/A	N/A	0.173	0.140	0.065	0.108
mCT 3	1.098	N/A	N/A	0.168	0.148	0.082	0.151
mCT 4	1.019	N/A	N/A	0.160	0.130	0.041	0.082
mCT 5	1.033	N/A	N/A	0.176	0.127	0.067	0.092
mCT 6	1.067	N/A	N/A	0.113	0.107	0.033	0.105
Discovery 710 1	1.139	N/A	N/A	0.151	0.176	0.051	0.188
Discovery 710 2	1.045	N/A	N/A	0.213	0.168	0.086	0.130
Discovery 710 1 Q.Clear 1	1.172	N/A	N/A	0.085	0.189	0.054	0.207
Discovery 710 2 Q.Clear 2	1.049	N/A	N/A	0.204	0.172	0.064	0.131
Discovery 710 3 Q.Clear 3	1.154	N/A	N/A	0.114	0.184	0.042	0.200
GE Discovery MI 1	1.055	N/A	N/A	0.105	0.100	0.032	0.095
GE Discovery MI 2	1.066	N/A	N/A	0.179	0.142	0.065	0.125
GE Discovery MI 1 Q.Clear 1	1.119	N/A	N/A	0.040	0.123	0.017	0.108
GE Discovery MI 2 Q.Clear 2	1.124	N/A	N/A	0.107	0.157	0.039	0.168
GE Discovery IQ 1	1.102	N/A	N/A	0.255	0.240	0.047	0.201
GE Discovery IQ 1 Q.Clear 1	1.083	N/A	N/A	0.234	0.219	0.052	0.177
Vereos 1	1.029	N/A	N/A	0.230	0.176	0.074	0.115
Min	0.911			0.040	0.100	0.017	0.082
Max	1.172			0.350	0.288	0.207	0.228
Average	1.063			0.193	0.180	0.078	0.148
COV _{MCR}	6.3%						

Table 11. Analysis results of 23 additional reconstructions using SUVmax quantitative metric.

Reconstruction mode	MCR	CRV _{medium}	CRV _{short}	Curvature	Absolute error	Curvature (excl. 10 mm sphere)	Absolute error (excl. 10 mm sphere)
Ingenuity 1	0.789	N/A	N/A	0.376	0.341	0.246	0.218
Ingenuity 2	0.736	N/A	N/A	0.405	0.383	0.284	0.267
mCT Flow 1	0.737	N/A	N/A	0.439	0.390	0.324	0.280
mCT Flow 2	0.750	N/A	N/A	0.476	0.379	0.353	0.263
mCT Flow 3	0.797	N/A	N/A	0.393	0.328	0.274	0.217
mCT 1	0.858	N/A	N/A	0.348	0.282	0.214	0.162
mCT 2	0.812	N/A	N/A	0.347	0.302	0.225	0.188
mCT 3	0.847	N/A	N/A	0.365	0.281	0.242	0.169
mCT 4	0.781	N/A	N/A	0.326	0.313	0.198	0.192
mCT 5	0.803	N/A	N/A	0.355	0.304	0.243	0.199
mCT 6	0.827	N/A	N/A	0.297	0.269	0.184	0.163
Discovery 710 1	0.829	N/A	N/A	0.357	0.301	0.234	0.188
Discovery 710 2	0.794	N/A	N/A	0.398	0.342	0.274	0.227
Discovery 710 1 Q.Clear 1	0.867	N/A	N/A	0.372	0.294	0.231	0.171
Discovery 710 2 Q.Clear 2	0.824	N/A	N/A	0.413	0.344	0.276	0.221
Discovery 710 3 Q.Clear 3	0.884	N/A	N/A	0.370	0.298	0.212	0.166
GE Discovery MI 1	0.797	N/A	N/A	0.351	0.313	0.233	0.202
GE Discovery MI 2	0.819	N/A	N/A	0.375	0.308	0.237	0.180
GE Discovery MI 1 Q.Clear 1	0.838	N/A	N/A	0.328	0.285	0.200	0.166
GE Discovery MI 2 Q.Clear 2	0.859	N/A	N/A	0.356	0.294	0.202	0.157
GE Discovery IQ 1	0.814	N/A	N/A	0.407	0.342	0.263	0.210
GE Discovery IQ 1 Q.Clear 1	0.831	N/A	N/A	0.412	0.336	0.258	0.199
Vereos 1	0.803	N/A	N/A	0.381	0.320	0.251	0.199
Min	0.736			0.297	0.269	0.184	0.157
Max	0.884			0.476	0.390	0.353	0.280
Average	0.813			0.376	0.320	0.246	0.200
COV _{MCR}	4.7%						

Table 12. Analysis results of 23 additional reconstructions using SUV_{peak} quantitative metric.

Sphere diameter (mm)	Current EARL RC bandwidth			Proposed RC bandwidth		
	SUVmean	SUVmax	SUVpeak	SUVmean	SUVmax	SUVpeak
37	0.76-0.89	0.95-1.16	N/A	0.85-1.00	1.05-1.29	0.99-1.07
28	0.72-0.85	0.91-1.13	N/A	0.82-0.97	1.01-1.26	0.95-1.07
22	0.63-0.78	0.83-1.09	N/A	0.80-0.99	1.01-1.32	0.90-1.09
17	0.57-0.73	0.73-1.01	N/A	0.76-0.97	1.00-1.38	0.75-0.99
13	0.44-0.60	0.59-0.85	N/A	0.63-0.86	0.85-1.22	0.45-0.69
10	0.27-0.43	0.34-0.57	N/A	0.39-0.61	0.52-0.88	0.27-0.41

Table 13. SUVmean, SUVmax and SUVpeak specifications proposed for harmonisation along with current EARL specifications.

Fig. 3. RC curves derived from additional reconstructions using SUVmean (a), SUVmax (b) and SUVpeak (c) quantitative metrics along with proposed new specifications. GE (Q.Clear) – blue dashed lines, GE (non-Q.Clear) – blue solid lines, Philips – red solid lines, Siemens – orange solid lines, possible new EARL specifications – black dashed lines.

3.4.7 Additional reconstructions

16 EARL accredited sites participated in the prospective evaluation of the newly proposed specifications for harmonisation and performed reconstructions according to instructions provided. Data received included 23 distinctive reconstructions from 3 GE Discovery 710 systems, 2 Philips Ingenuity systems, 6 Siemens mCT systems, 3 Siemens mCT Flow systems, 1 GE Discovery IQ system, 2 GE Discovery MI systems and 1 Philips Vereos system. RC curves derived from the 18 systems along with proposed new harmonising specifications can be seen in figure 3. For SUVmean, 16 out of 138 analysed spheres produced RC values outside of the suggested accreditation interval, while for SUVmax and SUVpeak, the number of outliers was 12. Quantitative results describing additional reconstructions can be found in Tables 10-12.

3.5 Discussion

The SUVmean and SUVmax RC curves of the initial 15 reconstruction modes vary significantly, even within one system. This reflects the high degree of variability that could be introduced into quantitative PET with variation in reconstruction settings. The selection of harmonising reconstruction modes and the validation which followed on additional reconstructions, demonstrated that the variability can be reduced to acceptable limits.

The acquisition time of 5 minutes per bed position specified in the current EARL accreditation settings, while characterising system performance in high statistics scenario, may not provide an accurate representation of the reconstruction mode's performance in clinical setting. Therefore, the observation of reduced CRV_{medium} and CRV_{short} in reconstruction modes for harmonisation is important since the acquisition times when utilising new PET/CT systems are routinely reduced to 2 minutes or less per bed position.

Significant increase in both SUVmean and SUVmax MCR values was observed in the reconstruction modes proposed for harmonisation compared to the corresponding current EARL specifications. The trend is in agreement with results recently published by Sunderland et al demonstrating that high-end PET/CT systems are having significantly increased SUVmax values in anthropomorphic phantom scans [53]. The metrics for all of the spheres demonstrated a noticeable increase, however for the smaller spheres (≤ 17 mm) the effect was relatively stronger. This could be explained by the so called Gibbs artefact which produces an overshoot of measured activity at the edges of the spheres, becoming more dominant at smaller sizes, also described by Lasnon et al [54]. To some extent the effect can be considered beneficial, compensating for the inherently lower recoveries seen in the smaller spheres.

It should however be noticed that with the use of resolution modelling (PSF) without any or with minimal post filtering applied, the overshoot could introduce significant positive SUV bias, in particular when using SUVmax. Methods like regularized (MAP) reconstruction with a regularizing prior (such as Q.Clear implemented by GE) can also be used to suppress Gibbs artefacts and were therefor also considered in this study.

The increased SUVmean and SUVmax recoveries seen in the proposed reconstruction modes for harmonisation would significantly reduce the gap that exists today between standardised quantitative reconstruction protocols used in multicentre settings and the locally developed non-standard protocols for lesion detection and general visual assessment – both of which are used in parallel in many nuclear medicine departments. Close agreement between the two could lead to the adoption of a single reconstruction mode that would provide standardised SUV data while maintaining increased lesion detectability.

In the reconstruction modes identified as suitable candidates for harmonisation, a relatively higher increase was found in the recoveries of smaller spheres. This would lead to more “flat” RC curves, making subsequent quantitative analysis less dependent on lesion size. With the proposed reconstruction modes, the recoveries remained largely size-independent for ≥ 17 mm diameter lesions. Moreover, it is important to notice that a possible new harmonising standard for systems with PSF implies SUVmax recoveries to exceed 1.0. This suggests that if SUVmax remains the *de facto* field standard for PET/CT quantification, one should accept a positive bias of about 10 to 25% for larger homogeneous objects (≥ 17 mm diameter).

For both SUVmean and SUVmax the proposed reconstruction modes for harmonisation yielded promising results. The two largest spheres (28 mm diameter, 37 mm diameter) showed excellent agreement across all systems for both SUVmean and SUVmax. Even though there isn't enough data for a reproducibility assessment, it can be predicted that a harmonising performance bandwidth is feasible for the next generation of PET/CT systems. The results from prospective validation using additional reconstructions will be further improved in the EARL accreditation process, where the centres will be guided to optimise their reconstruction settings in order to meet the new specifications.

As the harmonising RCs for SUVmean, SUVmax and SUVpeak all demonstrated a noticeable curve, the curvature and absolute error parameters exhibited increased or similar values with the initial reconstruction modes. Calculations excluding the smallest sphere demonstrated much better performance which illustrated the high impact the smallest sphere has, which led to a significant decrease in the RCs range.

The utility of SUVpeak was investigated as being a possible metric for standardised quantification. A recent prospective repeatability study by Kramer *et al.* [55] demonstrated

the robustness of using SUVpeak in non-small cell lung cancer patients. As previously shown by Makris *et al.* [56], and presented on supplemental figures 4-6, SUVpeak is significantly less sensitive to changes in reconstruction parameters and acquisition durations than SUVmean or SUVmax. The difference is mostly prominent in the initial group of 15 relatively loosely selected reconstruction modes, while within the 5 reconstructions for harmonisation and 23 additional ones, the difference became less apparent. On the other hand, the benefits of SUVpeak were offset by its consistently low recoveries for spheres with ≤ 17 mm diameter and therefore low MCR, which is comparable to that of SUVmean but significantly (20-40 %) lower than that of SUVmax. This is due to peak VOI size approaching or even exceeding the size of the sphere, therefore missing some of the active volume. If this issue could be addressed by, for example reducing the SUVpeak VOI size, SUVpeak may become an effective alternative to SUVmax, especially if quantitative comparison among reconstructions of unknown origin or non-harmonised PET/CT systems is desired. Harmonisation among systems remains necessary in order to enable reliable use of SUVmax. Further studies are needed in order to explore the optimal peak VOI diameter maintaining noise cancelling effects, while producing higher, yet harmonised recoveries.

An alternative to the described methodology of achieving harmonised recoveries, such as suggested in this paper, could be to gradually increase the post smoothing on high recovery PET data until harmonized RC-s are obtained (supplemental figures 9-11). Such a method is available on some systems and previously validated by Lasnon *et al* [54]. Potentially a post-smoothing feature on a workstation could be used for this purpose. This could result in higher recoveries and may reduce noise and Gibbs artefacts to acceptable levels for multicentre harmonisation. However, when offline post-smoothing needs to be applied to a dataset in order for it to achieve quantitative harmonisation, the filter information for the specific system always needs to accompany the PET data and extra care be taken that the filter be actually applied and clearly reported every time when required.

3.5.1 Limitations and future directions

Quantification of PET images is affected by uncertainties derived from reconstruction settings as well as global system (cross-) calibration. In this study the experimental data were corrected for global calibration errors, but in clinical practice both effects should be taken into consideration. Therefore, an accurate system calibration remains of utmost importance for all PET/CT systems used for quantification in order to keep the uncertainties as low as possible.

The phantom experiments conducted were sensitive to measurement uncertainties of dose calibrators and human error during the phantom preparation phase. The uncertainties related to phantom filling procedure are not part of this study and may increase the bandwidth of achievable harmonisation.

All experiments on various PET/CT models were performed on appointed systems. The inter-system variability stemming from the individual differences among the systems of same make and differences due to manufacturer allowed variability in well counter calibration factors, may increase the bandwidth of achievable harmonisation even further, although the newly proposed harmonisation specification were set using the same bandwidth as current EARL, which was shown to be appropriate and feasible.

As the position of VOI-s used in the analysis and comparison of SUVmean data is based on PET images rather than CT data, it is to some extent affected by image noise and may induce a small additional uncertainty to the results. This however is reflective of the clinically used method of VOI positioning. When this strategy is followed, it is therefore important to also put a threshold on acceptable noise levels (in this paper background noise should be lower than 15%). Yet, use of CT based VOI definition could be of interest in order to mitigate the effects of noise on VOI definition and subsequently on the measurement of the recovery coefficients. Another alternative could be the use of SUVpeak rather than SUVmax as starting point for VOI definition, as was applied in Frings *et al* [57]. These strategies may be considered when developing future standards.

Current study investigated harmonisation of PET/CT systems using ^{18}F tracer based FDG. The results cannot be directly transferred to system performance harmonisation involving other PET isotopes such as ^{68}Ga or ^{82}Rb which have a substantially longer positron range. System performance harmonisation with positron emitting isotopes other than ^{18}F requires further investigation.

In this feasibility study we primarily made use of reconstruction methods and parameter settings that were predefined or could be easily set by the user on commercially released systems. Where the software permitted, we applied additional reconstructions to include at least PSF and TOF, and also tried other reconstruction parameter settings which were expected to yield higher recoveries than current EARL specification. Yet, in this study we did not extensively explore a wide range of reconstruction settings as, e.g. iterations, subsets, matrix sizes etc., since our aim was to investigate clinically available protocols which are accessible for the users. Moreover, the investigated reconstruction modes had similar, but still different voxel sizes as well as the number of iterations/subsets between various systems which complicates direct comparison. In conclusion the harmonisation investigated in this study should be considered as a first feasibility test aiming at improving the current EARL specifications. Of course higher level of harmonisation would also be possible by considering more parameters, but then the question will be the feasibility in clinical practice. Further work is also needed to more extensively explore the impact of PSF reconstructions, voxel size and number of iterations/subsets on the variability of quantitative metrics of clinical datasets. Some reports have already been published showing that the repeatability and ICC of SUVmax, SUVpeak and SUVmean are at acceptable level [58].

To conclude, despite of possible limitations, we have studied the feasibility of harmonising state of the art PET/CT system performances and results suggest that an update of EARL current specification is feasible and achievable in practice.

3.6 Conclusions

This study investigated the feasibility of harmonising performance for PET/CT systems equipped with the latest Time-of-Flight (ToF) and resolution modelling (PSF) technology. Also, new possible specifications with higher contrast recoveries were investigated using various metrics such as average, maximum and peak SUV. Harmonising state of the art PET/CT systems with ToF and PSF technologies was found to be feasible. The harmonisation of such systems would require an update to the current multicentre accreditation program of EARL in order to accommodate higher recoveries. SUV_{peak} could be used as an uptake metric being less sensitive to noise and variation in image quality resulting from different reconstruction settings. It could be considered as an alternative to SUV_{max} if lower recoveries are considered to be acceptable for lesions of 17 mm diameter and smaller.

3.7 Acknowledgments

We would like to thank the EARL sites that provided additional reconstructions of the image quality QC phantom experiments, namely:

University Clinic for Nuclear Medicine and Endocrinology, Salzburg, Austria
 VU University Medical Centre, Amsterdam, The Netherlands
 GIE Humanitép, Lille, France
 University Hospital of Cologne, Cologne, Germany
 Ghent University Hospital, Ghent, Belgium
 Zentralklinik Bad Berka, Bad Berka, Germany
 University Medical Centre Utrecht, Utrecht, The Netherlands
 Centre Antoine Lacassagne, Nice, France
 VieCuri Medical Centre, Venlo, The Netherlands
 University Hospital Magdeburg, Magdeburg, Germany
 Cancer Institute Nantes-Atlantiques, St Herblain, France
 Groene Hart Ziekenhuis, Gouda, The Netherlands

University Hospital of Navarra, Pamplona, Spain
 University Hospital Antwerp, Edegem, Belgium
 Heinrich-Heine University Hospital, Düsseldorf, Germany
 University Hospital Brussels, Jette, Belgium
 Bispebjerg Hospital, Copenhagen, Denmark
 St Vincent's Hospital PET - CT Centre, Linz, Austria
 Postgraduate Institute of Medical Education and Research, Chandigarh, India
 King's College Hospital NHS Foundation Trust, London, United Kingdom
 Uppsala University Hospital, Uppsala, Sweden.

3.8 Disclaimer

This study and the authors do not represent the official position of the EANM Research Ltd (EARL). The proposed new accreditation specifications do not in any way replace the current standards but rather are a first evaluation on the feasibility of an update. The current EARL specifications are applicable until official notice from EANM Research Ltd.

3.9 Compliance with Ethical Standards

Andres Kaalep declares that he has no conflict of interest. Terez Sera has received travel grants and honoraria from EARL. Sjoerd Rijnsdorp declares that he has no conflict of interest. Maqsood Yaqub declares that he has no conflict of interest. Anne Talsma declares that he has no conflict of interest. Martin A. Lodge declares that he has no conflict of interest. Ronald Boellaard is an unpaid member of the EARL scientific advisory board.

This article does not contain any studies with human participants or animals performed by any of the authors.

3.10 Supplemental Material

	Siemens mCT	Siemens mCT Flow	GE 710	Philips Ingenuity
Sensitivity	10 kcps/MBq	10.2 kcps/MBq	7.5 kcps/MBq	7.3 kcps/MBq
Peak NECR	186 kcps @ 30.1 kBq/ml	180 kcps @ 28 kBq/ml	130 kcps @ 29.5 kBq/ml	124 kcps @ 2.3 kBq/ml
TOF resolution	540 ps	540 ps	N/A	540 ps
Crystal material	LSO	LSO	Lutetium based	LYSO
Crystal size	4 x 4 x 20 mm3	4 x 4 x 20 mm3	4.2 x 6.3 x 25 mm3	4 x 4 x 22 mm3
axial FOV	22.1 cm	22.1 cm	15.7 cm	18 cm
Resolution transverse 1cm	4.4 mm	4.3 mm	4.9 mm	N/A
Resolution axial 1cm	4.4 mm	4.3 mm	5.6 mm	N/A
Resolution transverse radial 10cm	5.7 mm	5.2 mm	N/A	N/A
Resolution transverse tangential 10cm	4.9 mm	4.7 mm	N/A	N/A
Axial 10 cm	5.9 mm	5.9 mm	6.3 mm	N/A

Supplemental table 1. Specifications of the four PET/CT systems used for the initial 15 reconstruction modes.

Parameter	GE 710, IQ & MI systems with Q.Clear option	GE 710, IQ & MI systems without Q.Clear	Philips Ingenuity TF	Siemens Bio-graph mCT Flow	Siemens Bio-graph mCT
Bed position duration (s)	300	300	300	0.5 mm/s (bed motion speed)	300
Recon method	QCFX, beta = 800	VPFXS	BLOB-OS-TF	TrueX-TOF (ultraHD-PET)	TrueX-TOF (ultraHD-PET)
Iterations/subsets	N/A	2/24	3/33	2 / 21	3 / 21
Matrix size	256x256	256x256	144x144	200x200	400x400
Post filter	N/A	5.0mm	N/A	Gaussian, 5.0mm	Gaussian, 5.0mm
Z-filter	Standard	Standard	N/A	N/A	N/A
PSF (Point Spread Function)	ON	ON	ON	ON	ON
ToF (Time of Flight)	ON*	ON	ON	ON	ON
Attenuation correction	ON, CT based	ON, CT based	ON, CT based	ON, CT based	ON, CT based
Scatter correction	ON	ON	ON	ON	ON

Supplemental table 2. Requested acquisition and reconstruction parameters during validation part with the 18 additional systems.

Supplemental Fig. 4. Variable sphere size (a to i) SUVmean recovery coefficients of Siemens, Philips and GE reconstructions plotted as a function of background noise (CoV). Reconstructions determined to be suitable for harmonisation are marked with triangles of the corresponding colour.

Supplemental Fig. 5. Variable sphere size (a to i) SUVmax recovery coefficients of Siemens, Philips and GE reconstructions plotted as a function of background noise (CoV). Reconstructions determined to be suitable for harmonisation are marked with triangles of the corresponding colour.

Supplemental Fig. 6. Variable sphere size (a to i) SUV_{peak} recovery coefficients of Siemens, Philips and GE reconstructions plotted as a function of background noise (CoV). Reconstructions determined to be suitable for harmonisation are marked with triangles of the corresponding colour.

Supplemental Fig. 7. Transversal slices from harmonising reconstructions a – Siemens 1 – D; b – Siemens 2 – A; c – Philips – B; d – GE – F and e – GE – E. Colour scale represents SUV values.

Supplemental Fig. 8. CoV_{BG} values for initial 15 reconstruction modes (a) and 5 proposed harmonising reconstruction modes (b).

Supplemental Fig. 9. RC curves derived from post-filtered Siemens 1 - A reconstruction using SUVmean (a), SUVmax (b) and SUVpeak (c) quantitative metrics along with proposed new EARL specifications.

Supplemental Fig. 10. RC curves derived from post-filtered GE - A reconstruction using SUVmean (a), SUVmax (b) and SUVpeak (c) quantitative metrics along with proposed new EARL specifications.

Supplemental Fig. 11. RC curves derived from post-filtered Philips - A reconstruction using SUVmean (a), SUVmax (b) and SUVpeak (c) quantitative metrics along with proposed new EARL specifications.

Bibliography

1. Avril NE, Weber WA. Monitoring response to treatment in patients utilizing PET. *Radiol. Clin. North Am.* 2005. p. 189–204.
2. E, Groen B, Jager PL, Cobben DCP, van der Graaf WTA, Vaalburg W, et al. The value of FDG-PET in the detection, grading and response to therapy of soft tissue and bone sarcomas; a systematic review and meta-analysis. *Cancer Treat. Rev.* 2004. p. 83–101.
3. Erdi YE. The use of PET for radiotherapy. *Curr. Med. Imaging Rev.* 2007;3:3–16.
4. Fletcher JW, Djulbegovic B, Soares HP, Siegel B a, Lowe VJ, Lyman GH, et al. Recommendations on the use of 18F-FDG PET in oncology. *J. Nucl. Med.* 2008;49:480–508.
5. de Geus-Oei L-F, van der Heijden HFM, Corstens FHM, Oyen WJG. Predictive and prognostic value of FDG-PET in nonsmall-cell lung cancer. *Cancer* [Internet]. Wiley Subscription Services, Inc., A Wiley Company; 2007;110:1654–64. Available from: <http://dx.doi.org/10.1002/cncr.22979>
6. Hoekstra CJ, Stroobants SG, Smit EF, Vansteenkiste J, van Tinteren H, Postmus PE, et al. Prognostic relevance of response evaluation using [18F]-2-fluoro-2-deoxy-D-glucose positron emission tomography in patients with locally advanced non-small-cell lung cancer. *J. Clin. Oncol.* 2005;23:8362–70.
7. Juweid ME, Stroobants S, Hoekstra OS, Mottaghy FM, Dietlein M, Guermazi A, et al. Use of positron emission tomography for response assessment of lymphoma: Consensus of the imaging subcommittee of international harmonization project in lymphoma. *J. Clin. Oncol.* 2007;25:571–8.
8. Vansteenkiste JF, Stroobants SG. The role of positron emission tomography with 18F-fluoro-2-deoxy-D-glucose in respiratory oncology. *Eur. Respir. J.* 2001;17:802–20.
9. Weber WA. Use of PET for Monitoring Cancer Therapy and for Predicting Outcome. *J. Nucl. Med.* [Internet]. 2005;46:983–95. Available from: <http://jnm.snmjournals.org/content/46/6/983.abstract>
10. Borst GR, Belderbos JS a, Boellaard R, Comans EFI, De Jaeger K, Lammertsma A a, et al. Standardised FDG uptake: a prognostic factor for inoperable non-small cell lung cancer. *Eur. J. Cancer.* 2005;41:1533–41.
11. Shankar LK, Hoffman JM, Bacharach S, Graham MM, Karp J, Lammertsma AA, et al. Consensus Recommendations for the Use of 18F-FDG PET as an Indicator of Therapeutic Response in Patients in National Cancer Institute Trials. *J Nucl Med.* 2006;47:1059–66.
12. Gupta T, Master Z, Kannan S, Agarwal JP, Ghosh-Laskar S, Rangarajan V, et al. Diagnostic performance of post-treatment FDG PET or FDG PET/CT imaging in head and neck cancer: a systematic review and meta-analysis. *Eur. J. Nucl. Med. Mol. Imaging.* 2011;38:2083.
13. Ung YC, Maziak DE, Vanderveen JA, Smith CA, Gulenchyn K, Lacchetti C, et al. 18Fluorodeoxyglucose positron emission tomography in the diagnosis and staging of lung cancer: A systematic review. *J. Natl. Cancer Inst.* 2007;99:1753–67.
14. Lv Y-L, Yuan D-M, Wang K, Miao X-H, Qian Q, Wei S-Z, et al. Diagnostic performance of integrated positron emission tomography/computed tomography for mediastinal lymph node staging in non-small cell lung cancer: a bivariate systematic review and meta-analysis. *J. Thorac. Oncol. International Association for the Study of Lung Cancer;* 2011;6:1350–8.
15. Eschmann SM, Friedel G, Paulsen F, Reimold M, Hehr T, Budach W, et al. 18F-FDG PET for assessment of therapy response and preoperative re-evaluation after neoadjuvant radio-chemotherapy in stage III non-small cell lung cancer. *Eur. J. Nucl. Med. Mol. Imaging.* 2007;34:463–71.
16. Hicks RJ, Kalff V, Macmanus MP, Ware RE, McKenzie AF, Matthews JP, et al. The Utility of F-FDG PET for Suspected Recurrent Non – Small Cell Lung Cancer After Potentially Curative Therapy : Impact on Management and Prognostic Stratification. 2001;42:10–21.
17. Dijkman BG, Schuurbiens OCJ, Vriens D, Looijen-Salamon M, Bussink J, Timmer-Bonte JNH, et al. The role of (18)F-FDG PET in the differentiation between lung metastases and synchronous second primary lung tumours. *Eur. J. Nucl. Med. Mol. Imaging. Berlin/Heidelberg: Springer-Verlag;* 2010;37:2037–47.
18. Hellwig D, Graeter TP, Ukena D, Groeschel A, Sybrecht GW, Schaefers HJ, et al. 18F-FDG PET for mediastinal staging of lung cancer:

- which SUV threshold makes sense? *J Nucl Med.* 2007;48:1761–6.
19. Agarwal M, Brahmanday G, Bajaj SK, Ravikrishnan KP, Wong C-YO. Revisiting the prognostic value of preoperative (18)F-fluoro-2-deoxyglucose ((18)F-FDG) positron emission tomography (PET) in early-stage (I & II) non-small cell lung cancers (NSCLC). *Eur. J. Nucl. Med. Mol. Imaging.* 2010;37:691–8.
 20. Liao S, Penney BC, Wroblewski K, Zhang H, Simon CA, Kampalath R, et al. Prognostic value of metabolic tumor burden on 18F-FDG PET in nonsurgical patients with non-small cell lung cancer. *Eur. J. Nucl. Med. Mol. Imaging.* 2012;39:27–38.
 21. Cheson BD, Pfister B, Juweid ME, Gascoyne RD, Specht L, Horning SJ, et al. Revised response criteria for malignant lymphoma. *J. Clin. Oncol.* 2007. p. 579–86.
 22. Hicks RJ. Role of 18F-FDG PET in Assessment of Response in Non-Small Cell Lung Cancer. *J. Nucl. Med.* 2009;50:31S–42.
 23. Czernin J, Weber WA, Herschman HR. Molecular imaging in the development of cancer therapeutics. *Annu. Rev. Med. United States;* 2006;57:99–118.
 24. Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. *Nat. Rev. Drug Discov. England;* 2003;2:566–80.
 25. Weber WA. Assessing Tumor Response to Therapy. *J Nucl Med.* 2009;50:1S–10.
 26. Weber WA, Petersen V, Schmidt B, Tyndale-Hines L, Link T, Peschel C, et al. Positron emission tomography in non-small-cell lung cancer: prediction of response to chemotherapy by quantitative assessment of glucose use. *J. Clin. Oncol. United States;* 2003;21:2651–7.
 27. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J. Nucl. Med.* 2009;50 Suppl 1:11S–20S.
 28. Boellaard R. Need for standardization of 18F-FDG PET/CT for treatment response assessments. *J. Nucl. Med. [Internet].* 2011;52 Suppl 2:93S–100S. Available from: http://jnm.snmjournals.org/content/52/Supplement_2/93S.full
 29. Doot RK, Pierce LA, Byrd D, Elston B, Allberg KC, Kinahan PE. Biases in Multicenter Longitudinal PET Standardized Uptake Value Measurements. *Transl. Oncol. [Internet].* 2014;7:48–54. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3998681&tool=pmcentrez&rendertype=abstract>
 30. Feuardent J, Soret M, De Dreuille O, Foehrenbach H, Buvat I. Reliability of uptake estimates in FDG PET as a function of acquisition and processing protocols using the CPET. *IEEE Trans. Nucl. Sci.* 2005;52:1447–52.
 31. Kurland BF, Gerstner ER, Mountz JM, Schwartz LH, Ryan CW, Graham MM, et al. Promise and pitfalls of quantitative imaging in oncology clinical trials. *Magn. Reson. Imaging.* 2012;30:1301–12.
 32. Lammertsma AA, Hoekstra CJ, Giaccone G, Hoekstra OS. How should we analyse FDG PET studies for monitoring tumour response? *Eur. J. Nucl. Med. Mol. Imaging.* 2006;33.
 33. Vriens D, Visser EP, De Geus-Oei LF, Oyen WJG. Methodological considerations in quantification of oncological FDG PET studies. *Eur. J. Nucl. Med. Mol. Imaging.* 2010. p. 1408–25.
 34. Westerterp M, Pruim J, Oyen W, Hoekstra O, Paans A, Visser E, et al. Quantification of FDG PET studies using standardised uptake values in multi-centre trials: Effects of image reconstruction, resolution and ROI definition parameters. *Eur. J. Nucl. Med. Mol. Imaging.* 2007;34:392–404.
 35. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur. J. Nucl. Med. Mol. Imaging.* 2014;42:328–54.
 36. Delbeke D, Coleman RE, Guiberteau MJ, Brown ML, Royal HD, Siegel B a, et al. Procedure guideline for tumor imaging with 18F-FDG PET/CT 1.0. *J. Nucl. Med.* 2006;47:885–95.
 37. Graham MM, Wahl RL, Hoffman JM, Yap JT, Sunderland JJ, Boellaard R, et al. Summary of the UPICT protocol for 18F-FDG PET/CT imaging in oncology clinical trials. *J. Nucl. Med.* 2015;56:955–61.
 38. Scheuermann JS, Saffer JR, Karp JS, Levering AM, Siegel A. Qualification of PET Scanners for Use in Multicenter Cancer Clinical Trials: The American College of Radiology Imaging Network Experience. *J. Nucl. Med.* 2010;50:1187–93.
 39. Kinahan P, Wahl R, Shao L, Frank R, Perlman E. The QIBA profile for quantitative FDG-PET/CT oncology imaging. *J. Nucl. Med.* 2014;55:1520–.
 40. Boellaard R, Willemsen A T, Arends B, Visser EP. EARL procedure for assessing PET/CT system specific patient FDG activity preparations for quantitative FDG PET/CT studies. 2013. p. 1–3.

41. Weber WA. PET for response assessment in oncology: Radiotherapy and chemotherapy. *Br. J. Radiol.* 2005;78:42–9.
42. Young H, Baum R, Cremerius U, Herholz K, Hoekstra O, Lammertsma AA, et al. Measurement of clinical and subclinical tumour response using [18F]- fluorodeoxyglucose and positron emission tomography: Review and 1999 EORTC recommendations. *Eur. J. Cancer.* 1999;35:1773–82.
43. Huang H. Anatomy of SUV. *Nucl Med Biol* [Internet]. 2000;27. Available from: [http://dx.doi.org/10.1016/S0969-8051\(00\)00155-4](http://dx.doi.org/10.1016/S0969-8051(00)00155-4)
44. Thie JA. Understanding the standardized uptake value, its methods, and implications for usage. *J. Nucl. Med.* 2004;45:1431–4.
45. Krak NC, Boellaard R, Hoekstra OS, Twisk JWR, Hoekstra CJ, Lammertsma AA. Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *Eur. J. Nucl. Med. Mol. Imaging.* 2005;32:294–301.
46. Sera T. Experiences and Results of the PET/CT Accreditation Program: The 150 Sites Milestone. *EANM Comm. Symp.* 2016.
47. Rausch I, Cal-González J, Dapra D, Gallowitsch HJ, Lind P, Beyer T, et al. Performance evaluation of the Biograph mCT Flow PET/CT system according to the NEMA NU2-2012 standard. *EJNMMI Phys.* 2015;2:26.
48. Karlberg AM, Sæther O, Eikenes L, Goa PE. Quantitative comparison of PET performance- Siemens Biograph mCT and mMR. *EJNMMI Phys.* [Internet]. 2016;3:5. Available from: <http://ejnmmiphs.springeropen.com/articles/10.1186/s40658-016-0142-7>
49. Kolthammer JA, Su K-H, Grover A, Narayanan M, Jordan DW, Muzic RF. Performance evaluation of the Ingenuity TF PET/CT scanner with a focus on high count-rate conditions. *Phys. Med. Biol.* [Internet]. 2014;59:3843–59. Available from: <http://iopscience.iop.org/article/10.1088/0031-9155/59/14/3843>
50. GE Healthcare Discovery PET/CT 710 system manufacturer webpage [Internet]. Available from: http://www3.gehealthcare.com/en/products/categories/pet-ct/pet-ct_scanners/discovery_pet-ct_710
51. Philips USA website [Internet]. Available from: <https://www.usa.philips.com/healthcare/product/HC882456/ingenuity-tf-pet-ct-scanner>
52. Martin A, Lodge1, Muhammad A. Chaudhry1, 2, and Richard L. Wahl1 3. Noise Considerations for PET Quantification Using Maximum and Peak Standardized Uptake Value. 2012;53:1041–7.
53. Sunderland JJ, Christian PE. Quantitative PET/CT Scanner Performance Characterization Based Upon the Society of Nuclear Medicine and Molecular Imaging Clinical Trials Network Oncology Clinical Simulator Phantom. *J Nucl Med.* 2015;56:145–52.
54. Lasnon C, Desmonts C, Quak E, Gervais R, Do P, Dubos-Arvis C, et al. Harmonizing SUVs in multicentre trials when using different generation PET systems: Prospective validation in non-small cell lung cancer patients. *Eur. J. Nucl. Med. Mol. Imaging.* 2013;40:985–96.
55. Kramer GM, Frings V, Hoetjes N, Hoekstra OS, Smit EF, de Langen AJ, et al. Repeatability of Quantitative Whole-Body 18F-FDG PET/CT Uptake Measures as Function of Uptake Interval and Lesion Selection in Non-Small Cell Lung Cancer Patients. *J. Nucl. Med.* [Internet]. 2016;57:1343–9. Available from: <http://jnm.snmjournals.org/cgi/doi/10.2967/jnumed.115.170225>
56. Makris NE, Huisman MC, Kinahan PE, Lammertsma AA, Boellaard R. Evaluation of strategies towards harmonization of FDG PET/CT studies in multicentre trials: Comparison of scanner validation phantoms and data analysis procedures. *Eur. J. Nucl. Med. Mol. Imaging.* 2013;40:1507–15.
57. Frings V, van Velden FHP, Velasquez LM, Hayes W, van de Ven PM, Hoekstra OS, et al. Repeatability of Metabolically Active Tumor Volume Measurements with FDG PET/CT in Advanced Gastrointestinal Malignancies: A Multicenter Study. *Radiology* [Internet]. Radiological Society of North America; 2014;273:539–48. Available from: <https://doi.org/10.1148/radiol.14132807>
58. van Velden FHP, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. *Mol. Imaging Biol.* [Internet]. Molecular Imaging and Biology; 2016;18:788–95. Available from: <http://dx.doi.org/10.1007/s11307-016-0940-2>

Chapter 4

Quantitative implications of the updated EARL 2019 PET-CT performance standards

Author(s): Andres Kaalep, Coreline N. Burggraaff, Simone Pieplenbosch,
Eline E. Verwer, Terez Sera, Josee Zijlstra, Otto S. Hoekstra,
Daniela E. Oprea-Lager, Ronald Boellaard

*as published in the European Journal of Nuclear Medicine and Molecular Imaging Physics
(Kaalep et al, 2019)*

4.1 Abstract

Purpose: Recently, updated EARL specifications (EARL2) have been developed and announced. This study aims at investigating the impact of the EARL2 specifications on the quantitative reads of clinical PET-CT studies and testing a method to enable the use of the EARL2 standards whilst still generating quantitative reads compliant with current EARL standards (EARL1).

Methods: Thirteen non-small cell lung cancer (NSCLC) and seventeen lymphoma PET-CT studies were used to derive four image datasets – the first dataset complying with EARL1 specifications and the second reconstructed using parameters as described in EARL2. For the third (EARL2F6) and fourth (EARL2F7) dataset in EARL2, respectively 6 mm and 7 mm Gaussian post filtering was applied. We compared the results of quantitative metrics (MATV, SUVmax, SUVpeak, SUVmean, TLG, tumor-to-liver and tumor-to-bloodpool ratio) obtained with these four datasets in 55 suspected malignant lesions using three commonly used segmentation/volume of interest (VOI) methods (MAX41, A50P, SUV4).

Results: We found that with EARL2 MAX41 VOI method, MATV decreases by 22%, TLG remains unchanged and SUV values increase by 23-30% depending on the specific metric used. The EARL2F7 dataset produced quantitative metrics best aligning with EARL1, with no significant differences between the most of the datasets ($p > 0.05$). Different VOI methods performed similarly with regards to SUV metrics but differences in MATV as well as TLG were observed. No significant difference between NSCLC and lymphoma cancer types was observed.

Conclusions: Application of EARL2 standards can result in higher SUVs, reduced MATV and slightly changed TLG values relative to EARL1. Applying a Gaussian filter to PET images reconstructed using EARL2 parameters, successfully yielded EARL1 compliant data.

Keywords: Performance, harmonisation, PET-CT, Quantification, EARL accreditation, standards.

4.2 Introduction

Positron emission tomography (PET) and computed tomography (CT) hybrid imaging (PET-CT) with ^{18}F -fluorodeoxyglucose (^{18}F -FDG) is widely being used in oncology for diagnosis, staging, restaging and therapy response evaluation due to its ability to measure metabolic changes [1–7]. In addition to visual inspection, quantitative PET data analysis [8] can provide additional benefits such as increased precision and reduced inter observer variability [9]. Standardised uptake value (SUV) is commonly used to represent the tissue radioactivity concentration normalised to the whole-body activity concentration, estimated from injected activity and body weight [10]. However, SUV bias and increased variability can arise from multiple factors [9,11] and need to be given extra consideration when multicentre data is desired or absolute quantitative measures used [9,12–15].

Therefore several scientific societies such as the European Association of Nuclear Medicine (EANM), American College of Radiology (ACR), American Association of Physicists in Medicine (AAPM), Radiological Society of North America (RSNA) and Society of Nuclear Medicine and Molecular Imaging (SNMMI) are promoting standardisation and harmonisation of imaging procedures and practices to reduce variability of quantification in a multicentre setting [16–20]. Results and experience with these programs are described in papers by Scheuermann et al. [19] Sunderland et al. [21] and Kaalep et al. [22].

In 2006, EANM Research Ltd. (EARL) initiative was launched by the EANM to promote multicentre nuclear medicine and research. In 2010, the EARL ^{18}F -FDG-PET-CT accreditation program was established to address variability in the quickly growing field of quantitative ^{18}F -FDG-PET imaging by setting up guidelines and specifications to which the participating sites must adhere. According to these guidelines, an accredited PET-CT system, in addition to other requirements, has to display a SUV bias of $\pm 10\%$ or less and produce contrast recoveries within a specified bandwidth (EARL1 and EARL2), when imaging hot spheres of various sizes within a NEMA NU2–2007 phantom.

The varying performance caused by multiple generations of PET-CT systems (2D, 3D acquisition, Time-of-Flight (TOF), etc.) and availability of various reconstruction technologies (e.g., resolution recovery/Point Spread Function (PSF) or Bayesian penalised-likelihood reconstruction) pose a particular problem when harmonisation within the community is desired. Multicentre standards should not be based on the least performing systems. They need to fit with the highest, yet common denominator in systems' performance.

The specifications for the EARL1 ^{18}F -FDG-PET-CT accreditation program were developed during a pilot study performed in 2010–2011, involving 12 PET-CT systems. Since then, the performance

of PET-CT systems has significantly increased and new acquisition and reconstruction technologies have been introduced [23]. A change in technology may also unfavourably affect patient management when previously used quantitative and visual criteria are applied without adaptation to the new images [24,25]. The harmonisation of these newer systems would require an update of the current multicentre accreditation standards (EARL1) in order to accommodate higher recoveries. A phantom study by Kaalep et al. [26], showed that the harmonisation of modern PET-CT systems from different vendors is feasible. Based on this study, the EARL1 specifications have been updated and EARL2 specifications were developed [26].

The introduction of an updated EARL standard changes quantitative PET-CT reads and these changes should be known and/or accounted for. Moreover, for ongoing multicenter studies it is advisable to continue generating data following the EARL1 standard to assure uniformity of image quantification for the entire multicenter dataset. The latter may be challenging for imaging sites as it could imply that 3 reconstructions are generated. One reconstruction following locally preferred settings optimised for lesion detectability, another reconstruction following EARL2 and a third reconstruction following EARL1 standards. Therefore, it is of interest to explore if an existing approach based on image filtering [27] can be applied to generate EARL1 compliant data from EARL2 reconstructed images. This would obviate the need to perform a (third) EARL1 compliant reconstruction. Moreover, it would still allow to generate both EARL2 and EARL1 compliant quantitative results to allow comparison of results with historic cohorts. Although in principle the image filtering approach can be applied to the clinically preferred reconstructions to generate either (or both) EARL1 or EARL2 compliant data, the filter settings required would vary from one site to another as locally preferred reconstruction settings are not the same. In a multicenter study this would require that these filter settings need to be derived, known and monitored for each system as this method does not generate EARL compliant images. Yet, deriving EARL1 compliant data from EARL2 reconstructed images is a more standardized condition or procedure and might be more easily reproduced elsewhere.

The primary aim of the current study is to investigate the impact of EARL2 updated accreditation specifications on the quantitative reads of clinical PET-CT studies. A secondary objective is to evaluate the performance of an (existing) approach based on image filtering to generate quantitative reads that are compliant with the EARL1 standards from EARL2 compliant reconstructed PET images.

4.3 Materials and methods

4.3.1 Selecting post-filtering parameters by phantom experiments

Twenty-one phantom images from a previously described study by Kaalep et al.[26] were investigated. These phantom images served to determine a post filter which, when applied to an EARL2 compliant dataset, would result in a dataset compliant with the EARL1 standards. The data were collected from 17 EARL accredited scanners from major vendors—3 Philips, 9 Siemens and 5 General Electric systems. The phantom experiments were performed in compliance with EARL Image Quality QC standard operating procedures. A NEMA NU2-2007 body phantom background was filled with a 2 kBq/ml ^{18}F -FDG solution and the 10, 13, 17, 22, 28 and 37 mm spheres with a 20 kBq/ml ^{18}F -FDG solution, resulting in a 10:1 sphere to background ratio when scanned for two 5 minute per bed positions. TOF, PSF, normalisation, randoms, scatter and attenuation corrections were applied. Reconstructions compliant with EARL1 as well as EARL2 specifications were performed. Reconstructed data were analysed using a semi-automatic tool developed for EARL [16]. Further details regarding the acquisition of the phantom data can be found in the initial study [26].

An additional Gaussian post-filtering with kernel sizes of 5, 6, 7 and 8 mm respectively, was applied to the EARL2 compliant phantom datasets using in-house post processing and analysis software ACCURATE [28]. Size dependent SUVmean and SUVmax recovery coefficients of all resulting datasets were compared with the EARL1 accreditation specifications in order to determine which filter values provided the largest number of EARL1 compliant results. A dataset was determined to be EARL1 compliant when observed contrast recoveries of SUVmean and SUVmax in all spheres were within EARL1 specifications. This strategy and methodology is equal to the one proposed and evaluated by Lasnon et al [27] and tested here to see if it could also be applied to derive EARL1 compliant results from EARL2 reconstructed PET data.

4.3.2 Patient selection and preparation

Thirty patients with non-small cell lung cancer (NSCLC, $n=13$) or lymphoma ($n=17$) were randomly selected from ongoing routine clinical staging or restaging studies with suspected positive lesions. The majority of lymphoma patients were diagnosed with diffuse large B-cell lymphoma. A standard uptake time of 60-75 min was applied for all patients. Further details can be found in Table 1.

Characteristic	Non-small cell lung cancer	Lymphoma	Total
Total no of patients	13	17	30
Males / Females	6 / 7	10 / 7	16 / 14
Weight of patients (mean)	42 – 92 (69.4) kg	61 – 103 (78.9) kg	42 – 103 (74.8) kg
BMI of patients (mean)	18 – 28 (23.3)	19 – 41 (25.7)	18 – 41 (24.7)
Administered activity (mean)	161 – 314 (240) MBq	205 – 347 (274) MBq	161 – 347 (259) MBq
Total no of analysed lesions	19	36	55
Lesions per patient (mean)	1-3 (1.5)	1-3 (2.1)	1-3 (1.8)

Table 1. Patient characteristics.

Acquisition and reconstruction parameters

30 patient scans performed on two EARL accredited Philips Ingenuity PET-CT TOF systems during a period of 2.5 years (from 10.03.2016 to 10.09.2018) were selected for further analysis within this study. For each patient study, two PET reconstructions were performed – first, using the EARL1 approved reconstruction parameters resulting in contrast recoveries within EARL1 accreditation specifications [29], and second, using the EARL2 compliant reconstruction parameters proposed by Kaalep et al [26]. Two additional image datasets were generated from the EARL2 reconstructed images by applying Gaussian filters of 6.0 mm (EAR2F6) and 7.0 mm (EAR2F7), respectively. Main parameters for the four PET image datasets used in this study are listed in Table 2.

Reconstruction	Pixel Spacing (mm)	Slice Thickness (mm)	Reconstruction Method	Post filter width (mm)	Resolution recovery
EARL1	4	4	BLOB-OS-TF	N/A	OFF
EARL2	4	4	BLOB-OS-TF	N/A	ON
EARL2F6	4	4	BLOB-OS-TF	6.0	ON
EARL2F7	4	4	BLOB-OS-TF	7.0	ON

Table 2. Main reconstruction parameters of the four investigated image datasets.

4.3.3 Lesion selection, segmentation and analysis

Lesions suspected to be malignant were identified by the author (AK) and confirmed by a certified nuclear medicine physician with 10 years of experience reading PET-CT images (DO). Physiological ^{18}F -FDG uptake (i.e. within brain, left ventricle, kidneys, urinary bladder) not related to the primary disease was excluded. A maximum of 3 lesions per patient was selected to avoid over-representation of any single patient, yielding a total of 55 lesions. VOIs were segmented semi-automatically using a region growing method with the following thresholds:

1. $\text{SUV} \geq 4.0$ (SUV_4)
2. 41% of SUV_{max} (41MAX)
3. 50% of SUV_{peak} with adaptation to local tumour-to-background contrast, so called adapted 50% of SUV_{peak} (A50P) [30]. SUV_{peak} is defined as the average uptake in a 1.2 cm diameter VOI positioned such to yield the highest value across all tumour voxels (also referred to as highest peak) [31].

Additionally for each patient, liver and bloodpool VOI-s were created and SUV_{max} , SUV_{peak} and SUV_{mean} calculated for EARL1, EARL2 and EARL2F7 reconstructions. For assessing liver uptake we positioned a 3 cm diameter spherical VOI in the right upper lobe of the (healthy) liver, as suggested by PERCIST [31]. For the blood pool uptake a 1.5 cm spherical VOI was positioned in the lumen of the ascending aorta.

For each combination of lesion, reconstruction and VOI threshold, the following quantitative metrics were calculated: metabolic active tumour volume (MATV), SUV_{mean} , SUV_{max} , SUV_{peak} and total lesion glycolysis (TLG) [30]. Moreover, we derived tumor-to-liver ratio's using SUV_{max} of both lesion and liver as well as SUV_{max} of the lesion and SUV_{mean} of the liver. Brief descriptions of the used VOI methods and quantitative metrics are given in Table 3.

VOI methods	Description
A50P	Region growing based VOI using 50% of SUV _{peak} with adaptation to local tumour-to-background contrast [32][30]
SUV4	Region growing based VOI using a threshold of SUV equal to 4
MAX41	Region growing based VOI using a threshold equal to 41% of SUV _{max}
Quantitative Metrics	Description
MATV	Volume of a lesion segmented using A50P, SUV4 or MAX41 method in mL
SUV _{mean}	Ratio of image derived average radioactivity concentration within a region of interest and the estimated whole body concentration of the injected radioactivity, normalised to bodyweight
SUV _{max}	Ratio of image derived maximum (single pixel) radioactivity concentration within a region of interest and the whole body concentration of the injected radioactivity, normalised to bodyweight
SUV _{peak}	Ratio of image derived average radioactivity concentration within a 12 mm diameter spherical volume (taking into account fractional voxels) within the region of interest, positioned to yield the highest uptake across all tumour voxel locations, and the whole body concentration of the injected radioactivity, normalised to bodyweight [30,31]
TLG	Total lesion glycolysis equal to the MATV times SUV _{mean}

Table 3. Descriptions of used VOI methods and quantitative metrics.

Statistical Analyses

Median relative differences of quantitative metrics determined from EARL2, EARL2F6 and EARL2F7 reconstructions and corresponding values from the EARL1 reconstruction were reported along with interquartile ranges. Mann-Whitney U-test and Wilcoxon signed-rank test were used to evaluate the statistical significance of the paired and non-paired data respectively.

4.4 Results

4.4.1 Phantom data

Post filtering values of 6 and 7 mm resulted in the highest number of EARL2 reconstructions (across all phantom scans) conforming to EARL1 specifications for both SUVmean and SUVmax (Fig. 1) and were selected for further analysis and testing using clinical datasets. As an example, Fig. 2 demonstrates the investigated systems' contrast recovery curves before and after the application of an additional 6 mm FWHM Gaussian post filter.

Fig. 1. Histogram of Gaussian post-filter values resulting in EARL1 compliant reconstructions of NEMA NU2-2007 body phantom data acquired in accordance with EARL guidelines for Image Quality QC standard operating procedures.

Fig. 2. Recovery coefficient values relative to sphere size determined from NEMA NU2-2007 body phantom data acquired in accordance with EARL guidelines for Image Quality QC standard operating procedures. **a** –SUVmean of EARL2 and EARL2F6 data; **b** – SUVmax of EARL2 and EARL2F6 data. Average measured EARL1 SUV data – solid black line, average measured EARL2F6 SUV data – solid red line current, standard deviation from average measured EARL1 SUV data – black dash dot dot line; standard deviation from average measured EARL2F6 SUV data – red dash dot dot line; EARL standard – bold black lines, prospective future EARL standard – bold red lines.

4.4.2 Clinical data

Figure 3 illustrates an example of a malignant lesion and the associated quantitative metrics calculated.

Fig. 3. Typical example image of a lymphoma patient's ^{18}F -FDG PET-CT scan. Coronal maximum intensity projection image is presented with an arrow pointing to a suspected malignant lesion along with the corresponding MATV, SUV_{max}, SUV_{peak}, SUV_{mean} and TLG quantitative metrics determined using EARL1, EARL2, EARL2F6 and EARL2F7 reconstructions.

We found that EARL2 SUV data were higher than EARL1 data ($p < 0.001$), the relative median difference ranging from 23% for SUVpeak to 25% for SUVmean and 30% for SUVmax (Table 4, Figs. 4c, 4d). The relative difference in SUV between the two datasets increased with decreasing MATV ($p < 0.001$) with a median difference of 36%, 39% and 25% (SUVmax, SUVmean and SUVpeak, respectively) for small (< 10 ml) lesions and 22%, 21% and 15% (SUVmax, SUVmean and SUVpeak, respectively) for large (≥ 10 ml) lesions (Fig. 4c), with no significant dependence on SUVmean i.e. lesion contrast (SUVmax, $p = 0.162$; SUVmean, $p = 0.225$; SUVpeak, $p = 0.178$) (Fig. 4d). SUV data from the filtered dataset EARL2F7 aligned best with those obtained using EARL1 reconstruction within -1% for SUVmax (interquartile range 9.5), -1% for SUVmean (interquartile range 9.4) and +2% for SUVmax (interquartile range 6.4).

Supplemental Fig. 4. Relative differences of 41MAX VOI method EARL2 and EARL2F7 reconstructions' MATV, SUVmax and TLG metrics compared to respective values from EARL1 reconstruction, presented as functions of EARL1 MATV and SUVmean. EARL2 reconstruction – red markers, EARL2F7 reconstruction – blue markers.

MATV seen on the EARL2 were on average 27% smaller ($p < 0.001$) when compared to EARL1 (Table 4, Figs. 4a, 4b), while the difference reduced to a statistically non-significant +2% ($p = 0.547$) when EARL2F7 and EARL1 were compared. The relative difference in EARL2 and EARL1 MATV values was found to be dependent on the underlying MATV values ($p < 0.001$) with an average difference of -31% for small (< 10 ml) lesions and -19% for large (≥ 10 ml) lesions (Fig. 4a). Significant MATV dependence ($p < 0.001$) remained when EARL2F7 and EARL1 were compared, with a median difference of +6% for small (< 10 ml) lesions and -8% for large (≥ 10 ml) lesions. A significant dependence of underlying SUV levels ($p = 0.033$) was found in MATV differences of EARL2 and EARL1, where lesions with $SUV_{mean} \leq 7.0$ demonstrated a median difference of -29% while lesions with $SUV_{mean} > 7.0$ demonstrated a median difference of -20%. Differences in MATV of EARL2F7 to EARL1 were independent ($p = 0.076$) on underlying SUV levels (Fig. 4b).

We found a statistically significant ($p = 0.005$) difference of -2% in median TLG values of EARL1 and EARL2 while no statistically significant difference in TLG derived from EARL1 and EARL2F7 ($p = 0.744$) or EARL1 and EARL2F6 ($p = 0.815$).

Figure 5 demonstrates the differences in SUV in liver and bloodpool VOIs between EARL1 and EARL2 and EARL2F7 reconstructions. The largest difference can be seen with SUV_{max} while SUV_{mean} demonstrates the smallest change. Figure 6 visualises tumor-to-liver ratio's calculated using various combinations of SUV metrics.

Supplemental Fig. 5. Relative differences of 41MAX VOI method EARL2 and EARL2F7 reconstructions' liver and blood pool SUVmean, SUVmax and SUVpeak metrics compared to respective values from EARL1 reconstruction, presented as functions of EARL1 SUVmean. EARL2 reconstruction – red markers, EARL2F7 reconstruction – blue markers.

Supplemental Fig. 6. Relative differences of 41MAX VOI method EARL2 and EARL2F7 reconstructions' tumor-to-liver ratios compared to respective values from EARL1 reconstruction, presented as functions of EARL1 MATV and SUVmean. **a** – tumor SUVmax to liver SUVmax ratio relative to MATV; **b** – tumor SUVmax to liver SUVmax ratio relative to SUVmean; **c** – tumor SUVmax to liver SUVmean ratio relative to MATV; **d** – tumor SUVmax to liver SUVmean ratio relative to SUVmean; **e** – tumor SUVpeak to liver SUVpeak ratio relative to MATV; **f** – tumor SUVpeak to liver SUVpeak ratio relative to SUVmean; **g** – tumor SUVmean to liver SUVmean ratio relative to MATV; **h** – tumor SUVmean to liver SUVmean ratio relative to SUVmean; EARL2 reconstruction – red markers, EARL2F7 reconstruction – blue markers.

Tumor-to-liver ratio was significantly higher for EARL2 when either SUVmax ($p < 0.001$) or SUVmean ($p < 0.001$) were used to measure liver SUV (Table 4). For EARL2F7 the difference with EARL1 data was still statistically significant ($p < 0.001$) when SUVmax was used to measure liver SUV and non-significant ($p = 0.344$) when liver SUVmean was used. All investigated metrics changed similarly between EARL2, EARL2F6 and EARL2F7 reconstructions regardless of cancer type (Fig. 7). SUVmax and SUVpeak metric changes remained similar while SUVmean, MATV and TLG behaviour differed based on the VOI method used (Supplemental Fig. 8). These differences between VOI methods were eliminated in the EARL2F7 reconstruction (Supplemental Fig. 8).

Cancer type	Reconstruction	MATV	SUVmax	SUVpeak	SUVmean	TLG	Tumor SUVmax / Liver SUVmax	Tumor SUVmax / Liver SUVmean
Lung cancer	EARL_V2	-22* (18)	30* (13)	23* (11)	25* (18)	-1 (8)	18* (10)	29* (12)
	EARL_V2F6	0 (8)	3* (6)	5* (3)	3* (5)	2 (5)	N/A	N/A
	EARL_V2F7	9 (13)	-3 (8)	1 (7)	-2 (7)	4 (12)	6* (9)	-3 (8)
Lymphoma	EARL_V2	-28* (18)	35* (18)	24* (12)	37* (19)	-3* (10)	26* (16)	35* (19)
	EARL_V2F6	-7* (12)	6* (11)	6* (7)	6* (12)	0 (8)	N/A	N/A
	EARL_V2F7	0 (13)	0 (9)	2 (6)	0 (9)	0 (10)	6* (7)	0 (9)
Both combined	EARL_V2	-27* (16)	33* (18)	23* (12)	34* (18)	-2* (8)	22* (15)	34* (17)
	EARL_V2F6	-5* (12)	5* (11)	5* (5)	5* (8)	1 (7)	N/A	N/A
	EARL_V2F7	2 (16)	-1 (10)	2* (6)	-1 (9)	2 (11)	6* (8)	1 (9)

Table 4. MAX41 VOI method relative median differences (%) of MATV, SUV metrics and TLG to corresponding values of EARL1 reconstruction along with corresponding interquartile ranges. Values marked with * indicate that the difference is statistically significant ($p < 0.05$).

Fig. 7. Comparison of relative differences of MATV (a), SUVmax (b), SUVmean (c), SUVpeak (d), TLG (e) and tumor-to-liver ratio (tumor SUVmax relative to liver SUVmean) (f) metrics between EARL1 and EARL2, EARL2F6, and EARL2F7 reconstructions, respectively, using 41MAX VOI method. Results obtained from lung cancer and lymphoma patients are presented separately. Central line of the box is the median, edges of the box are the 25th and 75th percentiles, the whiskers extend to either of the most extreme data points, which are not considered outliers or 1.5 times interquartile range. The outliers are marked using plus signs.

4.5 Discussion

Application of the updated EARL2 standards will affect quantitative reads. Yet, an update of the current EARL1 standards is required to cope with new PET-CT technologies providing enhanced lesion detectability. In this paper we studied the impact of using EARL2 standards on the quantitative reads of NSCLC and lymphoma ^{18}F -FDG PET-CT studies as compared to the EARL1 standards. An (existing) image filtering approach that enables the use of updated EARL2 compliant reconstruction, whilst still generating EARL1 compliant quantitative reads was derived and tested. The latter is important to allow comparison of data during ongoing studies and/or for comparison with historical cohorts.

As previously shown in the phantom study by Kaalep et al. [26], EARL2 compliant reconstruction resulted in a significant increase in contrast recovery and higher SUVs. The current study confirms these findings for clinical data (table 4) and demonstrates that overall trends (table 4, Fig. 7) were similar for both lung cancer and lymphoma patients, despite different lesion sizes and tracer uptake levels (or SUV). This suggests that the results of the current study could be universally applicable regardless of underlying disease type but this should be further investigated in future studies covering a wider range of patients and conditions.

Increase in contrast in EARL2 compliant reconstructions is similar for the most common used metrics SUVmax and SUVpeak applying all investigated VOI methods (Supplemental Fig. 8) and is independent of the lesion contrast (SUVmean). However, the results demonstrate a dependence on lesion volume where smaller lesions show a larger increase in contrast. The overall increased contrast recovery explains the wider and generally more preferred use of PSF reconstructions [23].

EARL2 compliant reconstructions result in a significant reduction in MATV values when MAX41 VOI method is used. Small (< 10 ml) lesions demonstrate a relatively higher decrease in volume than larger (≥ 10 ml) lesions. At the same time smaller lesions demonstrate larger increase in all SUV metrics which could be explained by the improved resolution and reduced spill-out effect caused by the PSF reconstruction. Post-filtering of the EARL2F6 and EARL2F7 reconstructions mimics the spill-out effect as the image is blurred by the filter resulting in loss of resolution and dispersion of measured activity in a larger volume. These results should not be transferred to other VOI methods as the PSF reconstruction affects different lesion segmentation methods differently.

TLG did not change significantly among the four investigated reconstructions. This may have been expected since while the lesion contrast (SUVmean) increases, MATV decreases proportionally, resulting in a reduced change in the product of the two. This reconstruction-independent behaviour of TLG could potentially make it a good metric for generating consistent quantitative measurements from both EARL1 and EARL2 standards without further image

modification. This is in agreement with Armstrong *et al.*, who suggested that TLG may be less sensitive to reconstruction methods compared with either SUVmax or SUVpeak [33]. TLG could also reduce the uncertainty of quantitative measurements of PET-CT scans that are performed on a system with unknown recovery coefficients. These results should not be transferred to SUV4 VOI methods as the PSF reconstruction affects the different lesion segmentations and therefore the corresponding TLG values also.

EARL1 compliant quantitative results can be generated from EARL2 reconstructed data using a simple filter. This approach has been applied and validated before by Lasnon *et al.*, specifically for PSF reconstruction on a Siemens Biograph PET-CT system using the so-called EQ.PET approach [27]. The method could however be considered limited due to lack of similar solutions by other major vendors. Moreover, the appropriate filter for each PET-CT system is not reported in the DICOM header and thereby the method cannot be applied offline or by 3rd party software. In our paper we identified (from phantom experiments) and verified the appropriate Gaussian filter setting in order to use EARL2 compliant reconstructions, while enabling the generation of EARL1 quantitative results. Also, it is important for verification of new datasets to be equivalent with historical cohorts. In this way data can be easily generated to conform both EARL1 and EARL2 standards without a need for additional reconstructions. For ongoing studies, however, it is recommended to keep adherence to the EARL1 standard and optionally adding EARL2 compliant reconstructions to gain understanding of quantitative implications when transitioning to the new standard.

It has been demonstrated by Kuhnert *et al.* that the increase in SUV, brought on by the use of PSF reconstruction, remains significant even after normalisation to the liver [24]. In our study, we also found that normalising lesion uptake to either liver SUVmax or liver SUVmean resulted in significantly increased tumor-to-liver ratio's and confirm the results from Kuhnert *et al.* that normalizing to liver uptake does not mitigate the effect of using different reconstructions on lesion uptake assessments. Apart from harmonizing quantitative ¹⁸F-FDG PET-CT reads, it has been reported that also visual assessment, e.g. Deauville scoring of PET-CT lymphoma studies, may be affected by a change in (reconstruction) technology and could affect patient management [25]. The observed changes (increases) in tumor-to-liver ratio's for EARL2 versus EARL1 compliant data suggest that use of EARL2 standards will result in overall higher Deauville scores, similarly as recently found by Ly *et al.* [34]. Therefore, it is not recommended to change standards and/or use new technologies without properly (re-)defining interpretation criteria. Hence use of EARL2 compliant reconstruction in combination with generating a second filtered dataset could be helpful in recalibrating these criteria for studies performed conform EARL1 performance standards to those obtained using updated performance standards. However, it should be noted that although this filtering approach has been shown to yield EARL1 compliant results and may facilitate these type of studies, this has not yet been demonstrated in this study nor was the aim.

4.6 Conclusions

Multicentre clinical trials require standardisation of quantitative results in order to be usable and exchangeable. This can be challenging as new acquisition and reconstruction technologies emerge and enable great benefits in image quality, but at the same time cause the quantitative performance to diverge.

In this paper we studied the impact of using updated EARL2 standards on the quantitative reads of NSCLC and lymphoma ^{18}F -FDG PET-CT studies as compared to the EARL1 standards. In general, the new EARL2 guidelines resulted in higher SUVs, smaller MATV and similar TLG values. A 7 mm FWHM Gaussian filter was shown to convert EARL2 compliant PET data to EARL1 compliant images. This facilitates the generation of both new and existing EARL compliant quantitative reads from a single EARL2 compliant image reconstruction.

4.7 Declarations

4.7.1 Abbreviations

NSCLC: non-small cell lung cancer; VOI: volume of interest; PET: positron emission tomography; CT: computed tomography; ^{18}F -FDG: ^{18}F -fluorodeoxyglucose; SUV: standardised uptake value; EANM: European Association of Nuclear Medicine; ACR: American College of Radiology; AAPM: American Association of Physicists in Medicine; RSNA: Radiological Society of North America; SNMMI: Society of Nuclear Medicine and Molecular Imaging; EARL: EANM Research Ltd.; TOF: time-of-flight; PSF: point spread function; BMI: body mass index; A50P: region growing based VOI using 50% of SUV_{peak} with adaptation to local tumour-to-background contrast; SUV₄: region growing based VOI using a threshold of SUV equal to 4; MAX₄₁: region growing based VOI using a threshold equal to 41% of SUV_{max}; MATV: volume of a lesion segmented using A50P, SUV₄ or MAX₄₁ method in mL; SUV_{mean}: ratio of image derived average radioactivity concentration within a region of interest and the estimated whole body concentration of the injected radioactivity, normalised to bodyweight; SUV_{max}: ratio of image derived maximum (single pixel) radioactivity concentration within a region of interest and the whole body concentration of the injected radioactivity, normalised to bodyweight; SUV_{peak}: ratio of image derived average radioactivity concentration within a 12 mm diameter spherical volume within the region of interest, positioned to yield the highest uptake, and the whole body concentration of the injected radioactivity, normalised to bodyweight; TLG: total lesion glycolysis equal to the MATV times SUV_{mean}.

4.7.2 Ethical Approval and Consent to participate

This retrospective study was approved by the Institutional Medical Ethics Committee (case number VUMC 2018.029), and the need for informed consent was waived. Consultation of the local objection registry verified that none of the selected patients had objected to use of their personal data for research purposes.

4.7.3 Consent for publication

Not applicable

4.7.4 Supplemental material

Supplemental Fig. 8. Comparison of relative differences of MATV (a), SUVmax (b), SUVmean (c), SUVpeak (d), TLG (e) and tumor-to-liver ratio (tumor SUVmax relative to liver SUVmean) (f) metrics between EARL1 and EARL2, EARL2F6, and EARL2F7 reconstructions, respectively, using different VOI methods. Central line of the box is the median, edges of the box are the 25th and 75th percentiles, the whiskers extend to either of the most extreme data points, which are not considered outliers or 1.5 times interquartile range. The outliers are marked using plus signs.

Supplemental Fig. 9. Relative differences of A50P VOI method EARL2 and EARL_V2F7 reconstructions' MATV, SUVmax and TLG metrics compared to respective values from EARL1 reconstruction, presented as functions of EARL1 MATV and SUVmean. EARL2 reconstruction – red markers, EARL2F7 reconstruction – blue markers.

Supplemental Fig. 10. Relative differences of SUV4 VOI method EARL2 and EARL2F7 reconstructions' MATV, SUVmax and TLG metrics compared to respective values from EARL1 reconstruction, presented as functions of EARL1 MATV and SUVmean. EARL2 reconstruction – red markers, EARL2F7 reconstruction – blue markers.

a - Difference in tumor-to-bloodpool ratio (tumor SUVmax / bloodpool SUVmax)
relative to MATV

b - Difference in tumor-to-bloodpool ratio (tumor SUVmax / bloodpool SUVmax)
relative to SUVmean

c - Difference in tumor-to-bloodpool ratio (tumor SUVmax / bloodpool SUVmean)
relative to MATV

d - Difference in tumor-to-bloodpool ratio (tumor SUVmax / bloodpool SUVmean)
relative to SUVmean

e - Difference in tumor-to-bloodpool ratio (tumor SUVpeak / bloodpool SUVpeak)
relative to MATV

f - Difference in tumor-to-bloodpool ratio (tumor SUVpeak / bloodpool SUVpeak)
relative to SUVmean

Supplemental Fig. 11. Relative differences of 41MAX VOI method EARL2 and EARL2F7 reconstructions' tumor-to-bloodpool ratio compared to respective values from EARL1 reconstruction, presented as functions of EARL1 MATV and SUVmean. **a** – tumor SUVmax to bloodpool SUVmax ratio relative to MATV; **b** – tumor SUVmax to bloodpool SUVmax ratio relative to SUVmean; **c** – tumor SUVmax to bloodpool SUVmean ratio relative to MATV; **d** – tumor SUVmax to bloodpool SUVmean ratio relative to SUVmean; **e** – tumor SUVpeak to bloodpool SUVpeak ratio relative to MATV; **f** – tumor SUVpeak to bloodpool SUVpeak ratio relative to SUVmean; **g** – tumor SUVmean to bloodpool SUVmean ratio relative to MATV; **h** – tumor SUVmean to bloodpool SUVmean ratio relative to SUVmean; EARL2 reconstruction – red markers, EARL2F7 reconstruction – blue markers

4.7.5 Competing interests

Andres Kaalep declares that he has no conflict of interest. Coreline N. Burggraaff declares that she has no conflict of interest. Simone Pieplensbosch declares that she has no conflict of interest. Eline E. Verwer declares that she has no conflict of interest. Terez Sera has received travel grants and honoraria from EARL. Josee Zijlstra declares that she has no conflict of interest. Otto S Hoekstra declares that he has no conflict of interest. Daniela E. Oprea-Lager declares that she has no conflict of interest. Ronald Boellaard is an unpaid member of the EARL scientific advisory board.

4.7.6 Funding

Not applicable

4.7.7 Authors' contributions

RB conceived of the study, participated in study design and coordination and helped to draft the manuscript. AK identified malignant lesions for further analysis, analysed the acquired data, participated in image processing and drafted the manuscript. CNB participated in the review of the paper. SP participated in the review of the paper. EV participated in identification and analysis of malignant lesions, image processing and review of the paper. TS participated in the review of the paper. JZ participated in the review of the paper. OSH participated in the review of the paper. DO verified the correct identification of malignant lesions and reviewed all cases. All authors approved the final manuscript.

4.7.8 Acknowledgements

Not applicable

Bibliography

1. Fletcher JW, Djulbegovic B, Soares HP, Siegel B a, Lowe VJ, Lyman GH, et al. Recommendations on the use of 18F-FDG PET in oncology. *J. Nucl. Med.* 2008;49:480–508.
2. Weber WA. Use of PET for Monitoring Cancer Therapy and for Predicting Outcome. *J. Nucl. Med.* 2005;46:983–95.
3. Borst GR, Belderbos JS a, Boellaard R, Comans EFI, De Jaeger K, Lammertsma A a, et al. Standardised FDG uptake: a prognostic factor for inoperable non-small cell lung cancer. *Eur. J. Cancer.* 2005;41:1533–41.
4. Shankar LK, Hoffman JM, Bacharach S, Graham MM, Karp J, Lammertsma AA, et al. Consensus Recommendations for the Use of 18F-FDG PET as an Indicator of Therapeutic Response in Patients in National Cancer Institute Trials. *J Nucl Med.* 2006;47:1059–66.
5. Gupta T, Master Z, Kannan S, Agarwal JP, Ghosh-Laskar S, Rangarajan V, et al. Diagnostic performance of post-treatment FDG PET or FDG PET/CT imaging in head and neck cancer: a systematic review and meta-analysis. *Eur. J. Nucl. Med. Mol. Imaging.* 2011;38:2083.
6. Ung YC, Maziak DE, Vanderveen JA, Smith CA, Gulenchyn K, Lacchetti C, et al. 18Fluorodeoxyglucose positron emission tomography in the diagnosis and staging of lung cancer: A systematic review. *J. Natl. Cancer Inst.* 2007;99:1753–67.
7. Liao S, Penney BC, Wroblewski K, Zhang H, Simon CA, Kampalath R, et al. Prognostic value of metabolic tumor burden on 18F-FDG PET in nonsurgical patients with non-small cell lung cancer. *Eur. J. Nucl. Med. Mol. Imaging.* 2012;39:27–38.
8. Weber WA. Assessing Tumor Response to Therapy. *J Nucl Med.* 2009;50:1S – 10.
9. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J. Nucl. Med.* 2009;50 Suppl 1:11S–20S.
10. Thie JA. Understanding the standardized uptake value, its methods, and implications for usage. *J. Nucl. Med.* 2004;45:1431–4.
11. Adams MC, Turkington TG, Wilson JM, Wong TZ. A systematic review of the factors affecting accuracy of SUV measurements. *AJR. Am. J. Roentgenol. United States;* 2010;195:310–20.
12. Fahey FH, Kinahan PE, Doot RK, Kocak M, Thurston H, Poussaint TY. Variability in PET quantitation within a multicenter consortium. *Med. Phys. United States;* 2010;37:3660–6.
13. Beyer T, Czernin J, Freudenberg LS. Variations in clinical PET/CT operations: results of an international survey of active PET/CT users. *J. Nucl. Med. United States;* 2011;52:303–10.
14. Boellaard R. Need for standardization of 18F-FDG PET/CT for treatment response assessments. *J. Nucl. Med. [Internet].* 2011;52 Suppl 2:93S–100S. Available from: http://jnm.snmjournals.org/content/52/Supplement_2/93S.full
15. Westerterp M, Pruim J, Oyen W, Hoekstra O, Paans A, Visser E, et al. Quantification of FDG PET studies using standardised uptake values in multi-centre trials: Effects of image reconstruction, resolution and ROI definition parameters. *Eur. J. Nucl. Med. Mol. Imaging.* 2007;34:392–404.
16. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur. J. Nucl. Med. Mol. Imaging.* 2014;42:328–54.
17. Delbeke D, Coleman RE, Guiberteau MJ, Brown ML, Royal HD, Siegel B a, et al. Procedure guideline for tumor imaging with 18F-FDG PET/CT 1.0. *J. Nucl. Med.* 2006;47:885–95.
18. Graham MM, Wahl RL, Hoffman JM, Yap JT, Sunderland JJ, Boellaard R, et al. Summary of the UPICT protocol for 18F-FDG PET/CT imaging in oncology clinical trials. *J. Nucl. Med.* 2015;56:955–61.
19. Scheuermann JS, Saffer JR, Karp JS, Levering AM, Siegel A. Qualification of PET Scanners for Use in Multicenter Cancer Clinical Trials: The American College of Radiology Imaging Network Experience. *J. Nucl. Med.* 2010;50:1187–93.
20. Kinahan P, Wahl R, Shao L, Frank R, Perlman E. The QIBA profile for quantitative FDG-PET/CT oncology imaging. *J. Nucl. Med.* 2014;55:1520–.
21. Sunderland JJ, Christian PE. Quantitative PET/CT Scanner Performance Characterization Based Upon the Society of Nuclear Medicine and Molecular Imaging Clinical Trials Network Oncology Clinical Simulator Phantom. *J Nucl Med.* 2015;56:145–52.
22. Kaalep A, Sera T, Oyen W, Krause BJ, Chiti A, Liu Y, et al. EANM/EARL FDG-PET/CT accreditation

- summary results from the first 200 accredited imaging systems. *Eur. J. Nucl. Med. Mol. Imaging. European Journal of Nuclear Medicine and Molecular Imaging*; 2018;45:412–22.
- 23. van der Vos CS, Koopman D, Rijnsdorp S, Arends AJ, Boellaard R, van Dalen JA, et al. Quantification, improvement, and harmonization of small lesion detection with state-of-the-art PET. *Eur. J. Nucl. Med. Mol. Imaging. European Journal of Nuclear Medicine and Molecular Imaging*; 2017;44:4–16.
- 24. Kuhnert G, Boellaard R, Sterzer S, Kahraman D, Scheffler M, Wolf J, et al. Impact of PET/CT image reconstruction methods and liver uptake normalization strategies on quantitative image analysis. *Eur. J. Nucl. Med. Mol. Imaging* [Internet]. 2016;43:249–58. Available from: <https://doi.org/10.1007/s00259-015-3165-8>
- 25. Barrington SF, Sulkin T, Forbes A, Johnson PWM. All that glitters is not gold - new reconstruction methods using Deauville criteria for patient reporting. *Eur. J. Nucl. Med. Mol. Imaging* [Internet]. 2018;45:316–7. Available from: <https://doi.org/10.1007/s00259-017-3893-z>
- 26. Kaalep A, Sera T, Rijnsdorp S, Yaqub M, Talsma A, Lodge MA, et al. Feasibility of state of the art PET/CT systems performance harmonisation. *Eur. J. Nucl. Med. Mol. Imaging. European Journal of Nuclear Medicine and Molecular Imaging*; 2018;45:1344–61.
- 27. Lasnon C, Salomon T, Desmonts C, Dô P, Oulkhovir Y, Madelaine J, et al. Generating harmonized SUV within the EANM EARL accreditation program: software approach versus EARL-compliant reconstruction. *Ann. Nucl. Med.* 2017;
- 28. Boellaard R. Quantitative oncology molecular analysis suite: ACCURATE. *J. Nucl. Med.* [Internet]. 2018;59:1753. Available from: http://jnm.snmjournals.org/content/59/supplement_1/1753.abstract
- 29. EANM EARL website [Internet]. Available from: <http://earl.eanm.org>
- 30. Frings V, van Velden FHP, Velasquez LM, Hayes W, van de Ven PM, Hoekstra OS, et al. Repeatability of Metabolically Active Tumor Volume Measurements with FDG PET/CT in Advanced Gastrointestinal Malignancies: A Multicenter Study. *Radiology* [Internet]. Radiological Society of North America; 2014;273:539–48. Available from: <https://doi.org/10.1148/radiol.14132807>
- 31. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. *J. Nucl. Med.* 2009.
- 32. Cheebsumon P, Yaqub M, Van Velden FHP, Hoekstra OS, Lammertsma AA, Boellaard R. Impact of [18F]FDG PET imaging parameters on automatic tumour delineation: Need for improved tumour delineation methodology. *Eur. J. Nucl. Med. Mol. Imaging.* 2011;
- 33. Armstrong IS, Kelly MD, Williams HA, Matthews JC. Impact of point spread function modelling and time of flight on FDG uptake measurements in lung lesions using alternative filtering strategies. *Arch. der Math.* 2014;
- 34. Ly J, Minarik D, Edenbrandt L, Wollmer P, Trägårdh E. The use of a proposed updated EARL harmonization of 18F-FDG PET-CT in patients with lymphoma yields significant differences in Deauville score compared with current EARL recommendations. *EJNMMI Res.* 2019;

Chapter 5

Feasibility of PET/CT system performance harmonisation for quantitative multicenter ^{89}Zr studies

Author(s): Andres Kaalep, Marc C. Huisman, Terez Sera,
Danielle Vugts, Ronald Boellaard, on behalf of EARL, EATRIS
and the TRISTAN Consortium (#IB4SD-116106)

*as published in the European Journal of Nuclear Medicine and Molecular Imaging Physics
(Kaalep et al, 2018)*

5.1 Abstract

Purpose: The aim of this study was to investigate the variability in quantitative performance and feasibility of quantitative harmonisation in ^{89}Zr PET/CT imaging.

Methods: Eight EANM EARL accredited [1] PET/CT systems were investigated using phantom acquisitions of uniform and NEMA NU2-2007 body phantoms. The phantoms were filled according to EANM EARL guidelines for [^{18}F]FDG but [^{18}F]FDG solution was replaced by a ^{89}Zr calibration mixture. For each system, standard uptake value (SUV) accuracy and recovery coefficients (RC) using SUVmean, SUVmax and SUVpeak metrics were determined.

Results: All eight investigated systems demonstrated similarly shaped RC curves and five of them exhibited closely aligning recoveries when SUV bias correction was applied. From the evaluated metrics, SUVpeak was found to be least sensitive to noise and reconstruction differences among different systems.

Conclusions: Harmonisation of PET/CT scanners for quantitative ^{89}Zr studies is feasible when proper scanner -dose calibrator cross-calibration and harmonised image reconstruction procedures are followed. An accreditation program for PET/CT scanners would facilitate multi-centre ^{89}Zr quantitative studies.

5.2 Introduction

The use of radiolabeled antibodies for diagnostic and therapeutic purposes has been going on for more than 50 years [2]. Their application as imaging probes in positron emission tomography (PET) combines the high sensitivity of PET with the high antigen specificity of monoclonal antibodies [3]. ^{89}Zr based tracers are becoming widespread with increasingly available supply, advances in radiochemistry and successful pilot studies in humans. However, multicentre studies using ^{18}F labelled tracers have demonstrated the need for standardisation of image acquisition, reconstruction, and analysis procedures and international harmonisation programs such as EANM and EARL aim to facilitate the use of FDG PET as a quantitative imaging biomarker [1,4]. A detailed discussion on ^{89}Zr physics in PET has been published by Conti et al [5].

The aim of this study was to investigate the variability in quantitative performance and feasibility of quantitative harmonisation in ^{89}Zr PET/CT imaging.

5.3 Materials and methods

5.3.1 Investigated systems & phantom experiments

Eight PET/CT systems (system 1 - 8), calibrated according to manufacturer's instructions, while also participating and accredited in the EANM/EARL [^{18}F]FDG PET/CT accreditation program, were selected for this study. The investigated systems were two General Electric Discovery 690; two General Electric Discovery 710; one Siemens Biograph 40 mCT; one Siemens Biograph 64 mCT; one Siemens Somatom Definition AS mCT and one Philips Ingenuity TF.

Two phantom experiments were carried out in accordance with EANM/EARL guidelines – Calibration QC and NEMA Phantom QC, where [^{18}F]FDG was substituted with a ^{89}Zr calibration sample. In the first experiment, a uniform cylindrical phantom was filled with a solution containing 8-12 kBq/mL of ^{89}Zr . In the second experiment the NEMA NU2-2007 body phantom background compartment and spheres were filled with a ^{89}Zr solution of 2 kBq/mL and 20 kBq/mL respectively, so as a 10:1 sphere to background ratio be achieved (Figure 1). Exact amount of ^{89}Zr activity was measured for each scan using only local dose calibrators, which had not undergone specific cross-calibration for ^{89}Zr . In both experiments the phantoms underwent a low-dose CT acquisition followed by PET acquisition of two consecutive bed positions of 5 minutes each. Images were reconstructed using EARL compliant parameters routinely used by the corresponding sites for [^{18}F]FDG quantitative imaging (Table 1).

Fig. 1. versal slice of ⁸⁹Zr filled NEMA NU2-2007 body phantom

System	System	Pixel Spacing (mm)	Slice Thickness (mm)	Reconstruction Method	Post filter width (mm)	Resolution recovery
A	Discovery 690	3.65	3.27	VPFXS	6.4	Yes
B	Discovery 690	2.73	3.27	VPHD	6.5	No
C	Discovery 710	2.73	3.27	VPFX	6.4	No
D	Discovery 710	2.73	3.27	VPFXS	9.0	Yes
E	Biograph 40 mCT	3.18	2.00	PSF+TOF	7.0	Yes
F	Somatom Definition AS mCT	3.18	5.00	PSF+TOF	8.0	Yes
G	Ingenuity TF PET/CT	4.00	4.00	BLOB-OS-TF	4.0	No
H	Biograph 64 mCT	4.07	5.00	OSEM3D+TOF	5.0	No

Table 1. Reconstruction settings

5.3.2 Data analysis

Reconstructed DICOM images were analysed using the EARL semi-automatic tool [1,6] designed for quantitative analysis of images of uniform and NEMA NU2-2007 body phantoms. From the uniform phantom and the NEMA body phantom's uniform background compartment, SUV accuracies for each system were determined. From the NEMA body phantom experiments, recovery coefficients (RC) were calculated as a function of sphere sizes, defined as ratio of activity concentration estimated from PET images to the expected activity concentration measured by dose calibrator. Different RC metrics values were calculated based on 50% background corrected isocontour VOI (SUVmean), maximum voxel value included in VOI (SUVmax) and spherical VOI with a diameter of 12 mm, positioned so as to yield the highest uptake (SUVpeak) [6][7][8]. Using data from the EARL database, relevant FDG RC curves of the corresponding scanners are displayed as a reference.

Additionally, RC curves were rescaled to correct for a global SUV bias, derived from the phantom's background compartment, to mitigate the impact of cross-calibration error between PET/CT system and dose calibrator on the observed RC. In order to directly compare the RC curves' shapes of all systems, the individual recovery coefficients of the NEMA body phantom spheres were normalised to the recovery coefficient of the largest (37 mm) sphere.

5.4 Results

The SUV bias from both phantom experiments are presented in Figure 2. The results for SUVmean, SUVmax and SUVpeak together with corresponding information for EARL [^{18}F]FDG can be seen in Figure 3, while the results corrected for SUV bias calculated from the body phantom background, are presented in Figure 4. Figure 5 demonstrates the RC curves normalised to the largest 37 mm sphere recovery.

Fig. 2. SUV accuracy of the PET scanners relative to the dose calibrator measurements plotted by system number. Systems 1-2 and 3-4 represent GE scanners Discovery 690 and Discovery 710, respectively; systems 5-6 are Siemens Biograph 40_mCT and Somatom Definition AS_mCT; system 7 is the Philips Ingenuity TF PET/CT; system 8 is the Siemens Biograph 64_mCT.

Fig. 3. RC curves derived from the ^{89}Zr phantom experiments using SUVmean (a), SUVmax (b) and SUVpeak (c) quantitative metrics and corresponding RC curves derived from the EARL ^{18}F FDG phantom experiments using SUVmean (d), SUVmax (e) and SUVpeak (f) quantitative metrics. Current EARL specifications for ^{18}F FDG-PET/CT accreditation are presented as bold dashed lines. Systems 1-2 and 3-4 represent GE scanners Discovery 690 and Discovery 710, respectively; systems 5-6 are Siemens Biograph 40_mCT and Somatom Definition AS_mCT; system 7 is the Philips Ingenuity TF PET/CT; system 8 is the Siemens Biograph 64_mCT.

Fig. 4. Background SUV bias corrected RC curves derived from the phantom experiments using SUVmean (a), SUVmax (b) and SUVpeak (c) quantitative metrics. Current EARL specifications for [^{18}F]FDG-PET/CT accreditation are presented as bold dashed lines. Systems 1-2 and 3-4 represent GE scanners Discovery 690 and Discovery 710, respectively; systems 5-6 are Siemens Biograph 40_mCT and Somatom Definition AS_mCT; system 7 is the Philips Ingenuity TF PET/CT; system 8 is the Siemens Biograph 64_mCT.

Fig. 5. RC curves normalised to the largest sphere, derived from the phantom experiments using SUVmean (a), SUVmax (b) and SUVpeak (c) quantitative metrics. Systems 1-2 and 3-4 represent GE scanners Discovery 690 and Discovery 710, respectively; systems 5-6 are Siemens Biograph 40_mCT and Somatom Definition AS_mCT; system 7 is the Philips Ingenuity TF PET/CT; system 8 is the Siemens Biograph 64_mCT.

5.5 Discussion

In order to remain in the optimal measurement range of the dose calibrators, the ^{89}Zr activity used in the study was similar to what is injected to a patient in clinical practice, resulting in significantly higher activity concentrations in the phantoms compared to patients (due to the smaller phantom volumes). However, lower counts are expected to further increase the variability of the results and may have hampered comparing recoveries between systems, with current EARL specifications and with those seen with ^{18}F . For clinical studies low count rates potentially induce an upward bias when SUVmax is used. To mitigate this upward bias, SUVpeak is an alternative less sensitive to scanner variation and image noise and might therefore be the optimal metric to assess tracer uptake for ^{89}Zr . Consequently, in our phantom study we included SUVpeak as well.

In addition to verifying the results of a recent study by Makris et al [9], current study investigated the real life scenario of using only local dose calibrators for ^{89}Zr measurement as well as by asking sites to perform the experiments themselves using the provided manuals and instructions. Out of the eight systems investigated in total, four Calibration QC and three NEMA Phantom QC experiments demonstrate a SUV bias of $>10\%$ (Figure 2). Since the scanners are EARL accredited for $[^{18}\text{F}]\text{FDG}$ -PET/CT, they comply with accreditation specifications for SUV bias ($\leq 10\%$), it is therefore believed that the large global errors are due to inaccurate cross-calibration between the scanners and dose calibrators used to measure the ^{89}Zr solution activity on site. While each of the dose calibrators should be set up by the manufacturer to accurately measure ^{89}Zr , the results from our study underline the importance of a traceable calibration performance of dose calibrators used in ^{89}Zr quantitative PET/CT imaging.

From Figure 2, it can be seen that SUV bias values derived from Calibration QC and NEMA Phantom QC background agree reasonably well, with the exception of only system 3 and to some extent system 4. These inconsistencies as well as the variable bias in RC curves (Figure 3) are suggested to be related to activity measurement and phantom filling procedures on site.

The initial RC curves derived from the images (Figure 3, a-c) demonstrate increased spread compared to the background corrected ones (Figure 4). After applying the SUV bias correction, the RC values of five systems show good alignment with each other and with EANM specifications for $[^{18}\text{F}]\text{FDG}$. Two of the investigated systems (1 and 7) remain out of specifications even after correcting for SUV bias. The reason for this is unknown and would need further investigation. RC curves normalised to the largest (37 mm) sphere (Figure 5) demonstrate similar shapes of RC curves for all investigated systems. This would suggest that, with further adjustment – meaning reduction of global SUV bias based on Calibration QC experiment data and possibly minor adjustment of the image reconstruction parameters – all of the systems should be able to achieve harmonisation.

The closest alignment of the RC curves can be observed when SUV_{peak} is used. This demonstrates the potential of this metric being used when quantitative harmonisation is desired. It should however be noted that with the use of SUV_{peak}, one should expect a decrease in overall contrast recovery, compared to SUV_{max}.

Finally, with the shape of ^{89}Zr RC curves shown to be similar to ^{18}F in our pilot study, further harmonisation efforts could be focused on the cross-calibration of the dose calibrators, which is considered to be the largest source of uncertainty in this case. A future ^{89}Zr harmonisation scheme could therefore be based on a ^{89}Zr dose calibrator cross-calibration quality control, with a successful site [^{18}F]FDG EARL accreditation being a prerequisite.

5.6 Conclusions

All eight investigated systems demonstrated similarly shaped RC curves and five of them exhibited close alignment when SUV bias correction was applied. Use of SUV_{peak} as a metric, which proved to be least sensitive to noise and reconstruction differences among systems, is strongly recommended for multicentre quantitative ^{89}Zr studies. When PET/CT and dose calibrator cross-calibration procedures are closely followed and the image reconstruction parameters adjusted, the quantitative harmonisation of scanners for ^{89}Zr PET studies is feasible. Yet, our results demonstrate the urgent need to set up a suitable cross-calibration and accreditation program to facilitate multi-centre ^{89}Zr quantitative studies.

5.7 Declarations

5.7.1 Ethics approval and consent to participate

This article does not contain any studies with human participants or animals performed by any of the authors.

5.7.2 Consent for publication

Not applicable.

5.7.3 Availability of data and materials

Data supporting the findings reported can be found at Zenodo:
<https://zenodo.org/record/1400583#.XVz4PeMzbAQ>

5.7.4 Competing Interest

Andres Kaalep declares that he has no conflict of interest. Marc Huisman declares that he has no conflict of interest. Terez Sera is a member of the EARL scientific advisory board and has received travel grants and honoraria from EARL. Danielle Vugts declares that she has no conflict of interest. Ronald Boellaard is an unpaid member of the EARL scientific advisory board.

5.7.5 Funding

The research leading to these results received funding from the Innovative Medicines Initiatives 2 Joint Undertaking under grant agreement No 116106. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

5.7.6 Authors contribution

RB conceived of the study, and participated in data collection, study design and coordination and helped to draft the manuscript. TS helped to draft the manuscript. AK analysed the acquired data and drafted the manuscript. MH participated in the data collection and analysis. DV participated in the data collection. All authors provided critical review and approved the final manuscript.

5.7.7 Acknowledgments

We would like to thank the EARL sites that provided data for the experiments, namely:

University Hospital Olomouc, Olomouc, Czech Republic

Academisch Ziekenhuis Groeninge Hospital, Kortrijk, Belgium

University College London, London, United Kingdom

Turku PET Centre, Turku, Finland

VU University Medical Centre, Amsterdam, The Netherlands

Radboud University Medical Center, Nijmegen, The Netherlands

University Hospital of Navarra, Pamplona, Spain

Bibliography

1. Kaalep A, Sera T, Oyen W, Krause BJ, Chiti A, Liu Y, et al. EANM/EARL FDG-PET/CT accreditation - summary results from the first 200 accredited imaging systems. *Eur. J. Nucl. Med. Mol. Imaging*. 2017;1–11.
2. McCardle RJ, Harper P V, Spar IL, Bale WF, Andros G, Jiminez F. Studies with Iodine-131-Labeled Antibody to Human Fibrinogen for Diagnosis and Therapy of Tumors. *J. Nucl. Med.* [Internet]. 1966;7:837–47. Available from: <http://jnm.snmjournals.org/content/7/11/837.short>
3. Zhang Y, Hong H, Cai W. PET Tracers Based on Zirconium-89. *Curr. Radiopharm.* 2011;4:131–9.
4. Aide N, Lasnon C, Veit-Haibach P, Sera T, Sattler B, Boellaard R. EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. *Eur. J. Nucl. Med. Mol. Imaging*. 2017.
5. Conti M, Eriksson L. Physics of pure and non-pure positron emitters for PET: A review and a discussion. *EJNMMI Phys.* [Internet]. EJNMMI Physics; 2016;3. Available from: <http://dx.doi.org/10.1186/s40658-016-0144-5>
6. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur. J. Nucl. Med. Mol. Imaging*. 2014;42:328–54.
7. Kinahan P, Wahl R, Shao L, Frank R, Perlman E. The QIBA profile for quantitative FDG-PET/CT oncology imaging. *J. Nucl. Med.* 2014;55:1520–.
8. Martin A, Lodge1, Muhammad A. Chaudhry1, 2, and Richard L. Wahl1 3. Noise Considerations for PET Quantification Using Maximum and Peak Standardized Uptake Value. 2012;53:1041–7.
9. Makris NE, Boellaard R, Visser EP, de Jong JR, Vanderlinden B, Wiertz R, et al. Multicenter Harmonization of 89Zr PET/CT Performance. *J. Nucl. Med.* [Internet]. 2014;55:264–7. Available from: <http://jnm.snmjournals.org/cgi/doi/10.2967/jnumed.113.130112>

Chapter 6

Future perspectives and concluding remarks

6.1 Future directions

A sensible future direction for any quantitative multicentre accreditation would be the extension of existing quality control regimens beyond the widely used [^{18}F]FDG to include emerging novel isotopes and tracers. The application of ^{68}Ga is increasing because of affordable production by using a $^{68}\text{Ge}/^{68}\text{Ga}$ generator and the possibility to label both small compounds and macromolecules [1]. Due to their slower pharmacokinetics, using monoclonal antibodies as therapeutic target-specific agents has generated interest in radioisotopes with longer half-lives, such as ^{64}Cu , ^{76}Br , ^{89}Y , ^{89}Zr and ^{124}I [2][3][4]. These have also demonstrated potential in theranostics - molecular imaging in tandem with a therapeutic agent, which could be either pharmaceutical or radiotherapeutic. Imaging with these isotopes however brings new challenges to quantification such as emitted and scattered prompt gammas within the energy window of the PET scanner, high energy gammas with associated pair production and high energy positrons escaping patient, among others [5]. Quantification of pure positron emitters such as ^{89}Zr is relatively simple and possibly could be tied to an existing ^{18}F quantification scheme. In this way, if the PET scanners are harmonised for ^{18}F , then further harmonisation to ^{89}Zr would require minimal additional efforts – possibly in the form of an additional global scanner calibration control. Quantification of ^{76}Br , ^{86}Y , ^{82}Rb and ^{124}I is more complicated since the random coincidences from prompt gamma contamination could blur or add a non-uniform background to the images. In the case of non-pure positron emitters, the feasibility of quantification depends on better randoms or prompt gamma correction methods to be developed by the PET system vendors.

Future research into advancing quantitative PET imaging will involve the review, extension of existing and inclusion of additional quality control metrics and tools to further improve the accuracy and consistency of quantitative reads. As PET systems advance, the accuracy of activity concentration estimation increases. This will facilitate for state-of art PET devices the application of more strict accreditation and standardisation criteria than currently in place, but, as a result some of the older systems will not comply with the more stringent requirements [6]. Additional quality control tools are under development, which may include the use of image derived quality metrics – extraction of quality metrics directly and preferably seamlessly from routine clinical studies [7,8]. For instance, general scanner calibration could be checked by evaluating the total amount of activity detected within the study field of view. This however requires further investigation into tracer distributions in clinical studies and the development and validation of a quality control tools for this purpose.

Standardisation programs for various imaging tasks and modalities such as quantitative brain imaging of Alzheimer's patients, gated cardiac PET or PET/MR could be devised to take into account the specific technical and task related features of possible failure and therefore tailor quality metrics and ways of standardisation in order to reduce the quantitative variability among different imaging sites and various imaging systems.

Recently, whole body parametric PET imaging has become a reality, allowing to generate full quantitative parametric images for the whole body [9]. A discussion is ongoing on whether this will replace or be added to the commonly performed procedure of multibed acquisitions, resulting in a static PET image from where conventional semi quantitative image derived PET metrics, such as the Standardized Uptake Value (SUV), are then derived and used in the context of clinical oncology [10]. Using Patlak analysis, the acquisition of functional parameters such as metabolic rate might even be achieved by making only two static late scans while the input function can be measured directly from PET images. Additional investigation is required to determine the accuracy and repeatability of quantitative data obtained by this method and whether the current standardisation programs will suffice for whole body dynamic PET investigations or new dedicated schemes will have to be developed.

Emergence of total body PET systems, i.e. systems with extended axial field of view (up to 2 m) such as in the EXPLORER PET/CT project, opens up new possibilities, but, poses difficulties as well, regarding the quality control procedures for such devices. Quality control phantoms currently in use cannot cover the entire field of view of a whole body scanner, which is a prerequisite in many quality control measurements. Therefore, either a longer phantom has to be developed or a technique worked out on how a shorter phantom is to be moved through the scanners field of view and account for each axial location of the system.

Radiomics is another rapidly developing field which provides large amounts of quantitative data based on image features such as texture, surface, shape and position of objects, surrounding tissue among many others, which may uncover disease characteristics unseen by conventional review of images [11][12]. These data are extracted from images using specific data characterisation algorithms which themselves may require further standardisation of more specific image quality features not in the attention of the human reader. It may also require the development of specific non-spherical and/or heterogeneous phantom inserts to measure radiomics features using more realistic tumour uptake distributions from those that currently can be achieved using the typical spherical and uniform inserts [13].

Finally, artificial intelligence (AI) and machine learning is rapidly making its way to nuclear medicine as well as to other kinds of medical imaging. AI can mine data (including image data) for patterns and anomalies and then use them to make predictions and recommendations to the reporting physician [14]. Further research is required to assess how much consistency in images is required for AI to work most effectively and the need for harmonisation programs specifically designed to facilitate AI applications. However, at present this field is still fully in development and it is too early to provide recommendations at this stage. Yet, initial studies show that standardisation of the imaging procedures and equipment performances are needed for using radiomics/AI analysis in multicentre studies as well [15].

6.2 Concluding remarks

In recent years PET has become a widespread functional and quantitative imaging tool. Due to its quantitative nature and being widely used in multicentre trials, it is essential to have quality assurance and quality control procedures in place to ensure the accuracy and precision of quantitation. In this thesis, the experience and results of running a large scale EANM-EARL [^{18}F]FDG PET/CT accreditation program is described and analysed. An update, taking into account the latest advances in PET/CT systems such as time-of-flight and resolution modelling, to the existing EARL accreditation criteria is investigated and developed. The impact of the newly developed and updated EARL harmonising program on quantitative reads of clinical PET/CT studies is investigated and a method enabling the use of updated standards while still generating quantitative reads, compliant with existing EARL standards, is tested. Additionally, in order to explore the possibility to expand the accreditation program to additional tracers, the feasibility of ^{89}Zr PET/CT quantitative harmonisation is demonstrated.

Bibliography

1. Velikyan I. Prospective of ^{68}Ga -Radiopharmaceutical development. *Theranostics*. 2014.
2. Reichert J. Monoclonal Antibodies as Innovative Therapeutics. *Curr. Pharm. Biotechnol.* 2008;
3. Vallabhajosula S, Solnes L, Vallabhajosula B. A broad overview of positron emission tomography radiopharmaceuticals and clinical applications: What is new? *Semin. Nucl. Med.* 2011.
4. Belov V V., Bonab AA, Fischman AJ, Heartlein M, Calias P, Papisov MI. Iodine-124 as a label for pharmacological PET imaging. *Mol. Pharm.* 2011;
5. Conti M, Eriksson L. Physics of pure and non-pure positron emitters for PET: A review and a discussion. *EJNMMI Phys.* [Internet]. *EJNMMI Physics*; 2016;3. Available from: <http://dx.doi.org/10.1186/s40658-016-0144-5>
6. Kaalep A, Sera T, Oyen W, Krause BJ, Chiti A, Liu Y, et al. EANM/EARL FDG-PET/CT accreditation - summary results from the first 200 accredited imaging systems. *Eur. J. Nucl. Med. Mol. Imaging. European Journal of Nuclear Medicine and Molecular Imaging*; 2018;45:412–22.
7. Boktor RR, Walker G, Stacey R, Gledhill S, Pitman AG. Reference range for intrapatient variability in blood-pool and liver SUV for ^{18}F -FDG PET. *J. Nucl. Med.* 2013;
8. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. *J. Nucl. Med.* 2009.
9. Karakatsanis NA, Lodge MA, Tahari AK, Zhou Y, Wahl RL, Rahmim A. Dynamic whole-body PET parametric imaging: I. Concept, acquisition protocol optimization and clinical application. *Phys. Med. Biol.* 2013;
10. Leahy R, Boellaard R, Zaidi H. Whole-body parametric PET imaging will replace conventional image-derived PET metrics in clinical oncology. *Med. Phys.* 2018;
11. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RGP, Granton P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer.* 2012;
12. Yip SSF, Liu Y, Parmar C, Li Q, Liu S, Qu F, et al. Associations between radiologist-defined semantic and automatically computed radiomic features in non-small cell lung cancer. *Sci. Rep.* 2017;
13. Pfaehler E, Beukinga RJ, de Jong JR, Slart RHJA, Slump CH, Dierckx RAJO, et al. Repeatability of ^{18}F -FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. *Med. Phys.* 2019;
14. Hall M. Artificial intelligence and nuclear medicine. *Nucl. Med. Commun.* [Internet]. 2018/10/26. Lippincott Williams & Wilkins; 2019;40:1–2. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/30362987>
15. Boellaard R. Limitations and Challenges Of AI in Standardization in Both Imaging and Analysis. *Artif. Intell. Methods Appl. Radiol. Nucl. Med.* 2019.

Chapter 7

Summary of findings

In recent years positron emission tomography has become a widespread functional imaging tool used for diagnosing, staging and restaging of various malignancies. As a quantitative imaging method, it is mandatory to have quality assurance and quality control procedures in place so as to ensure the accuracy and precision of quantitation. Additionally, the number of multicentre trials involving PET, where the requirement is to have uniform quantitative performance across all PET/CT systems, increases and quantitative evaluation and endpoints are more frequently used [1][2]. A particular problem to uniform quantitative imaging is posed by different levels of technology in imaging centres [3]. This requires the standardisation of imaging procedures and harmonisation of PET/CT system performance to achieve proper multicentre performance. EANM-EARL is running a PET/CT accreditation program which aims at facilitating multicentre studies based on the EANM Guidelines [4] for standardisation of [^{18}F]FDG PET/CT imaging and is harmonising the system performances by applying a self-developed procedure.

Chapter 2 describes the six years' experience of running the EANM-EARL [^{18}F]FDG PET/CT accreditation programme based on the first set of EARL standards and reports the findings and impact on the harmonised PET/CT systems. During the reported period, more than 2500 quality control datasets were collected from 200 systems in 150 imaging sites worldwide. Over the course of their accreditation period, each participating centre was requested to submit scans of two different phantoms - calibration quality control (CalQC), using a uniform cylindrical phantom and image quality control (IQQC), using the NEMA NU2-2007 body phantom. Average volumetric SUV bias and SUV recovery coefficients (RC) were calculated by EARL using a semi-automatic analysis program, then the data evaluated on the basis of quality control type, approval status, PET/CT system manufacturer and submission order. It was shown that SUV bias in 5% ($n = 96$) of all CalQC submissions ($n = 1816$) exceeded the acceptance level (10%). After corrective actions following communication of the results, the accredited sites achieved 100% compliance with EARL specifications. 30% ($n = 1381$) of SUVmean and 23% ($n = 1095$) of SUVmax sphere recoveries from first IQQC submissions failed to meet EARL accreditation criteria. After harmonisation of the scanners, for the subsequent submissions the failure rate decreased to 12% ($n = 360$) and 9% ($n = 254$), respectively. Most systems demonstrated longitudinal SUV bias reproducibility within $\pm 5\%$, while RC values remained stable and generally within $\pm 10\%$ for the four largest and $\pm 20\%$ for the two smallest spheres. It can be concluded that regardless of manufacturer or model, all investigated systems were able to comply with the EARL specifications. Within the EARL accreditation program, gross PET/CT calibration errors are successfully identified and longitudinal variability in PET/CT performances reduced. The program demonstrates that a harmonising accreditation procedure is feasible and achievable.

Based on the standardisation experience, **Chapter 3** explores the possibility of modernising EARL FDG accreditation programme by harmonising the performance of state of the art PET/CT systems equipped with time-of-flight (ToF) and resolution modelling/point spread function (PSF) technologies. Additionally, a method helping the new harmonising criteria with

higher contrast recoveries to comply with current EARL standards using various SUV metrics is developed. For the experiment, four PET/CT systems with both ToF and PSF capabilities from three major vendors were used. NEMA NU2–2007 body phantom scans were acquired and analysed using the EANM-EARL accreditation procedures. A total of 15 reconstruction parameter sets of varying pixel size, post filtering and reconstruction type, with three different acquisition times were used to compare the quantitative performance of the systems. A target range for recovery curves was established which accommodated the highest matching recoveries from all investigated systems. These updated criteria were validated on 18 additional scanners from 16 EARL accredited sites in order to demonstrate the scanners' ability to meet the new target range. Each of the four systems was found to be capable of producing harmonising reconstructions with similar recovery curves. The reconstruction parameter sets producing harmonising results, significantly increased SUVmean (25%) and SUVmax (26%) contrast recoveries compared with current EARL specifications. The additional prospective validation performed on 18 scanners from 16 EARL accredited sites demonstrated the feasibility of updated harmonising specifications. SUVpeak was found to significantly reduce the variability in quantitative results while producing lower recoveries in smaller (≤ 17 mm diameter) sphere sizes. Harmonising PET/CT systems with ToF and PSF technologies from different vendors was found to be feasible. The harmonisation of such systems would require an update of the current multicentre EARL accreditation program in order to accommodate higher recoveries. SUVpeak should be further investigated as a noise resistant alternative quantitative metric to SUVmax.

In **Chapter 4** the impact of the newly developed and updated EARL harmonising criteria on quantitative reads of clinical PET/CT studies is investigated and a method enabling the use of updated standards, while still generating quantitative reads compliant with existing EARL standards, is tested. To this end, thirteen non-small cell lung cancer (NSCLC) and seventeen lymphoma PET/CT studies were used to derive four image datasets – the first dataset (EARL1) complying with existing EARL specifications and the second (EARL2) reconstructed using parameters complying with new proposed EARL specifications. For the third (EARL2F6) and fourth (EARL2F7) datasets, respectively 6 mm and 7 mm Gaussian post filtering was applied to the EARL2 dataset. We compared the results of quantitative metrics (MATV, SUVmax, SUVpeak, SUVmean, TLG) obtained with these four datasets in 55 suspected malignant lesions using three commonly used segmentation/volume of interest (VOI) methods (MAX41, A50P, SUV4). It was found that with EARL2 MAX41 VOI method, MATV decreases by 25%, TLG remains unchanged and SUV values increase by 23–32% depending on the specific metric used. The EARL2F7 dataset produced quantitative metrics best aligning with EARL1, with no significant differences between the datasets ($p > 0.05$). Different VOI methods performed similarly with regards to SUV metrics but differences in MATV as well as TLG were observed. No significant difference between NSCLC and lymphoma cancer types was observed. Application of EARL2 standards can result in higher SUVs, reduced MATV and unchanged TLG values relative to EARL1. Applying a Gaussian filter to PET images reconstructed using EARL2 parameters, successfully yielded EARL1 compliant data.

Chapter 5 expands beyond FDG and investigates the variability in quantitative performance and feasibility of quantitative harmonisation in ^{89}Zr PET/CT imaging. Eight EANM-EARL PET/CT systems accredited for ^{18}F FDG were investigated using scans from uniform and NEMA NU2-2007 body phantoms. The phantoms were filled conform to the EANM-EARL accreditation specifications for ^{18}F FDG but the ^{18}F FDG solution was replaced by an ^{89}Zr calibration mixture. For each system, standard uptake value (SUV) accuracy and recovery coefficients (RC) using SUVmean, SUVmax and SUVpeak metrics were determined. All eight investigated systems demonstrated similarly shaped RC curves and five of them exhibited closely aligning recoveries when SUV bias correction was applied. From the evaluated metrics, SUVpeak was found to be the least sensitive to noise and reconstruction differences among different systems. It can be concluded that harmonisation of PET/CT scanners for quantitative ^{89}Zr studies is feasible when proper scanner to activity meter cross-calibration and harmonised image reconstruction procedures are followed. An accreditation program for ^{89}Zr PET/CT imaging would facilitate multi-centre ^{89}Zr quantitative studies.

Nederlandse samenvatting

Positron emissie tomografie (PET) is de laatste jaren een wijdverspreide functionele medisch beeldvormende techniek geworden voor de diagnose, stadiëring en her-stadiëring van verschillende maligniteiten. Voor het gebruik als kwantitatieve techniek is het noodzakelijk een kwaliteitssysteem te hebben voor de borging en controle van de nauwkeurigheid en precisie van de kwantitatieve resultaten. Bovendien neemt het aantal multicenter onderzoeken waarbinnen kwantitatieve PET wordt toegepast toe. Kwantitatieve evaluatie en kwantitatieve eindpunten worden steeds vaker toegepast, waarbij een uniforme basis voor harmonisatie van kwantitatieve resultaten een voorwaarde is [1][2]. De aanwezigheid van verschillende niveaus met betrekking tot de stand der techniek tussen de verschillende deelnemende instituten vormt een uitdaging voor het verkrijgen van uniforme en vergelijkbare resultaten [3]. EANM-EARL is een PET/CT accreditatie programma met het doel om multicenter studies te faciliteren door standaardisering, via de EANM Guidelines [4], van protocollen en de harmonisatie van de kwantitatieve prestaties van de gebruikte PET/CT systemen.

Hoofdstuk 2 beschrijft de zes jaar ervaring met het uitvoeren van het EANM-EARL [¹⁸F] FDG PET/CT accreditatieprogramma op basis van de eerste set EARL normen en rapporteert de bevindingen en impact daarvan op de kwantitatieve prestaties van de geharmoniseerde PET/CT systemen. Tijdens de gerapporteerde periode werden meer dan 2500 datasets voor kwaliteitscontrole verzameld van 200 systemen op 150 locaties wereldwijd. In de loop van hun accreditatieperiode werd elk deelnemend centrum verzocht om scans van twee verschillende fantomen in te dienen: kalibratie-kwaliteitscontrole (CalQC), met behulp van een uniform cilindrisch fantoom, en beeld-kwaliteitscontrole (IQQC), met behulp van het NEMA NU2-2007 fantoom. De gemiddelde SUV afwijking en SUV recoverycoëfficiënten (RC) werden berekend door EARL met behulp van een semi-automatisch analyseprogramma en vervolgens werden de gegevens geëvalueerd op basis van het soort kwaliteitscontrole, goedkeuringsstatus, PET/CT fabrikant en indieningsvolgorde. Er werd aangetoond dat de SUV afwijking in 5% ($n = 96$) van alle CalQC-inzendingen ($n = 1816$) het acceptatieniveau (10%) overschreed. Na corrigerende maatregelen na mededeling van de resultaten, waren 100% van de geaccrediteerde sites in staat te voldoen aan de EARL specificaties. Verder was gevonden dat 30% ($n = 1381$) van de gemiddelde SUV waarden en 23% ($n = 1095$) van maximale SUV waarden voor de bollen in het NEMA fantoom van eerste IQQC inzendingen niet voldeden aan de EARL accreditatiecriteria. Na harmonisatie van de scanners daalde het aantal afwijkingen voor de daarvop volgende inzendingen tot respectievelijk 12% ($n = 360$) en 9% ($n = 254$). De meeste systemen vertoonden een lange termijn SUV voorspelbaarheid binnen $\pm 5\%$, terwijl RC waarden in het algemeen stabiel bleven binnen $\pm 10\%$ voor de vier grootste en $\pm 20\%$ voor de twee kleinste bollen. Geconcludeerd kan worden dat, ongeacht de fabrikant of het model van de PET/CT camera, alle onderzochte systemen konden voldoen aan de EARL specificaties. Binnen het EARL

accreditatieprogramma worden grove PET/CT kalibratiefouten met succes geïdentificeerd en lange termijn variabiliteit in PET/CT prestaties verminderd. Het programma toont aan dat een harmoniserende accreditatieprocedure haalbaar is.

Gebaseerd op de ervaring met standaardisatie, wordt **hoofdstuk 3** het onderzoek beschreven naar de mogelijkheid om het EARL FDG accreditatieprogramma te moderniseren door de prestaties van geavanceerde PET/CT systemen, uitgerust met time-of-flight (ToF) en/of resolutiemodellering/point spread-functie (PSF) technologieën, te harmoniseren. Bovendien is een methode ontwikkeld die de nieuwe harmonisatiecriteria voor SUV met hogere recovery waarden helpt te voldoen aan de huidige EARL normen. Voor het experiment werden vier PET/CT systemen met zowel ToF als PSF mogelijkheden van drie grote leveranciers gebruikt. Scans van het NEMA NU2-2007 fantoom werden verkregen en geanalyseerd met behulp van de EANM-EARL accreditatieprocedures. Een totaal van 15 reconstructieparametersets met variërende pixelgrootte, filtering en reconstructiealgoritme en met drie verschillende acquisitietijden werden gebruikt om de kwantitatieve prestaties van de systemen te vergelijken. Een doelbereik voor recovery curves werd vastgesteld zodanig dat deze overeenkomst tussen resultaten maximaliseerde. Deze bijgewerkte criteria werden gevalideerd op 18 extra scanners van 16 EARL geaccrediteerde sites om aan te tonen dat de scanners in staat zijn om aan het nieuwe doelbereik te voldoen. Elk van de vier systemen bleek in staat te zijn data te produceren met vergelijkbare recovery curves. De nieuw reconstructieparametersets produceren harmoniserende resultaten met aanzienlijk hogere SUVmean (25%) en SUVmax (26%) recoveries in vergelijking met de huidige EARL specificaties. De aanvullende prospectieve validatie uitgevoerd op 18 scanners van 16 EARL geaccrediteerde sites toonde de haalbaarheid aan van de bijgewerkte specificaties. SUVpeak bleek de variabiliteit in kwantitatieve resultaten aanzienlijk te verminderen en tegelijkertijd lagere recovery te produceren in kleinere (≤ 17 mm diameter) bollen. Harmonisatie van PET/CT systemen met ToF en PSF technologieën van verschillende leveranciers bleek haalbaar. De harmonisatie van dergelijke systemen zou een update van het huidige multicenter EARL accreditatieprogramma vereisen om hogere SUV recoveries mogelijk te maken. SUVpeak moet verder worden onderzocht als een meer ruisbestendig alternatieve kwantitatieve maat in plaats van SUVmax.

In **hoofdstuk 4** wordt de impact van de nieuw ontwikkelde en bijgewerkte EARL harmonisatiecriteria op kwantitatieve metingen van klinische PET/CT onderzoeken bestudeerd en wordt een methode getest die het gebruik van de bijgewerkte normen mogelijk maakt, terwijl nog steeds kwantitatieve metingen worden gegenereerd die voldoen aan bestaande EARL normen. Daartoe werden dertien niet-kleincellige longkanker (NSCLC) en zeventien lymfoom PET/CT onderzoeken gebruikt om vier beelddatasets af te leiden - de eerste dataset (EARL1) die voldoet aan de bestaande EARL-specificaties en de tweede (EARL2) gereconstrueerd met behulp van parameters die voldoen aan de nieuwe voorgestelde EARL specificaties. Voor de derde (EARL2F6) en vierde (EARL2F7) gegevenssets werd respectievelijk 6 mm en 7 mm

Gaussische postfiltering toegepast op de EARL2 PET beelden. We vergeleken de resultaten van kwantitatieve maten (MATV, SUVmax, SUVpeak, SUVmean, TLG) verkregen met deze vier datasets van 55 vermoedelijke kwaadaardige laesies met behulp van drie veelgebruikte segmentatie / volume of interest (VOI) methoden (MAX41, A50P, SUV4). Het bleek dat met de EARL2 MAX41 VOI methode de MATV met 25% afneemt, de TLG ongewijzigd blijft en de SUV-waarden met 23-32% stijgen, afhankelijk van de gebruikte maat. De EARL2 F7 dataset produceerde kwantitatieve resultaten die het beste overeenkomen met EARL1, zonder significante verschillen tussen de datasets ($p > 0,05$). Verschillende VOI methoden presteerden vergelijkbaar met betrekking tot de diverse bekeken SUV maten, maar verschillen in zowel MATV als TLG werden waargenomen. Er werd geen significant verschil tussen NSCLC- en lymfoomkankertypen waargenomen. Toepassing van EARL2 normen kan leiden tot hogere SUV's, lager MATV en ongewijzigde TLG waarden ten opzichte van EARL1. Het toepassen van een Gaussisch filter op PET beelden gereconstrueerd met behulp van EARL2 parameters, leverde gegevens op die overeenkwamen met die verkregen met EARL1 parameters.

Hoofdstuk 5 gaat verder dan FDG en onderzoekt de variabiliteit in kwantitatieve prestaties en haalbaarheid van kwantitatieve harmonisatie in PET/CT beeldvorming met ^{89}Zr . Acht EANM-EARL PET/CT systemen geaccrediteerd voor $[^{18}\text{F}]\text{FDG}$ werden onderzocht met behulp van scans van het uniforme en het NEMA NU2-2007 fantoom. De fantomen werden gevuld conform de EANM-EARL accreditatiespecificaties voor $[^{18}\text{F}]\text{FDG}$ maar de $[^{18}\text{F}]\text{FDG}$ oplossing werd vervangen door een kalibratiemengsel van ^{89}Zr . Voor elk systeem werden de standaard uptake value (SUV) nauwkeurigheid en recoverycoëfficiënten (RC) met behulp van SUVmean, SUVmax en SUVpeak bepaald. Alle acht onderzochte systemen vertoonden gelijke RC curves en vijf van hen vertoonden dicht bij elkaar liggende recoveries wanneer SUV afwijkingscorrectie werd toegepast. Uit de geëvalueerde meetwaarden bleek SUVpeak het minst gevoelig voor ruis- en reconstructieverschillen tussen verschillende systemen. Geconcludeerd kan worden dat een harmonisatie van PET/CT systemen voor kwantitatieve ^{89}Zr onderzoeken haalbaar is wanneer de juiste kruiskalibratie van scanner, dosiskalibrator wordt uitgevoerd en geharmoniseerde beeldreconstructieprocedures worden gevolgd. Een accreditatieprogramma voor ^{89}Zr PET/CT beeldvorming zou multicenter ^{89}Zr kwantitatieve studies vergemakkelijken.

Bibliography

1. Doot RK, Kurland BF, Kinahan PE, Mankoff DA. Design Considerations for using PET as a Response Measure in Single Site and Multicenter Clinical Trials. *Acad. Radiol.* 2012;
2. Yankeelov TE, Mankoff DA, Schwartz LH, Lieberman FS, Buatti JM, Mountz JM, et al. Quantitative imaging in cancer clinical trials. *Clin. Cancer Res.* 2016.
3. Sunderland JJ, Christian PE. Quantitative PET/CT Scanner Performance Characterization Based Upon the Society of Nuclear Medicine and Molecular Imaging Clinical Trials Network Oncology Clinical Simulator Phantom. *J Nucl Med.* 2015;56:145–52.
4. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur. J. Nucl. Med. Mol. Imaging.* 2014;42:328–54.

Curriculum Vitae

Andres Kaalep was born in Tallinn, Estonia on the 31st of July in 1985. In 2007, he received a BSc in Engineering Physics in Tallinn University of Technology. In 2010, he completed an MSc in Engineering Physics in Tallinn University of Technology, with a specialization in applied physics. From 2006, Andres has worked as a biomedical engineer in North Estonia Medical Centre and from 2011, as the head of department of Imaging diagnostics and radiation therapy equipment. From 2013, he has been teaching master level courses in biomedical engineering in Tallinn Technical University. From 2016, he has been a PhD student in Faculty of Medical Sciences in The University of Groningen. The results of his doctoral research are presented in this thesis.

Acknowledgements

As I started down this journey some five years ago, I had no idea that it would lead me here. I have always wished to give something back to the scientific community and society at whole for schooling and raising me the level that I would be able to. One could consider it a moral obligation but I see it rather as a pleasant possibility to stretch one's mental capabilities and hopefully do something useful at the same time. What started out as an interesting side project next to my day to day work has suddenly evolved into this finalised PhD thesis you are now holding in your hands. This great achievement is not something that would have seen the light of day without the many great people around me whose support and wisdom have been invaluable.

First of all I would like to thank my supervisor prof. R. Boellaard, who has been instrumental in guiding me through this all from the start. Dear Ronald, your immense experience and wisdom have guided me while your always positive and encouraging attitude have not left any room for doubt that I would make it to the end. It has been a great honour and pleasure to work alongside you for all these years.

I would like to wholeheartedly thank both of my co-supervisors and especially prof. T. Sera. Without you Terez, I would not be where I am today. Thank you for introducing me to the EARL team and for always keeping track of my progress while critically reviewing all of my papers and providing invaluable feedback. I value your support greatly and hope that our future collaboration will be as fruitful as it has been so far.

I would like to thank the members of the assessment committee, prof. J. Pruim, Prof. S. Stroobants and Prof. H.W.A.M. De Jong, who have found time in their busy schedules to read my thesis and participate in the ceremony.

I would like to thank all of the co-authors of my papers and everybody who participated in different ways that made this work possible. Many people were important for this thesis to come into being and I would like to extend my gratitude towards everybody that I have forgot to mention but have had a role to play in this work.

Last but not least, I would like to thank my fiancée Annika, who has stood by and supported me tirelessly and has never held against me the long nights and early mornings that I spent behind the computer with my PhD instead of you. I will do my best to make it up to you nonetheless.