Invited lecture

Objectives and current status of QIBA (Quantitative Imaging Biomarkers Alliance)*1

Daniel C. Sullivan *2
Department of Radiology, Duke University Medical Center

Abstract

A quantitative imaging biomarker (QIB) is an objectively measured characteristic derived from an in vivo image as an indicator of normal biological processes, pathogenic processes or response to a therapeutic intervention. In 2007 the Radiological Society of North America (RSNA) organized the Quantitative Imaging Biomarkers Alliance (QIBA) whose mission is to improve the value and practicality of quantitative imaging biomarkers by reducing variability across devices, patients and time.

The QIBA initiative involves: (1) stakeholder collaboration to identify needs and solutions to develop consistent and reliable quantitative imaging results across imaging platforms, clinical sites, and time to achieve accurate and reproducible quantitative results from imaging methods. Since the process of acquiring a clinical imaging scan is complex, the goal requires much coordinated work among many stakeholders.

There are several sources of variability in quantitative results from clinical images: (1) image acquisition hardware, software and procedures; (2) measurement methods; and (3) reader variability. QIBA employs a consensus-driven approach to produce a QIBA Profile that includes one of more QIBA Claims and specifications for the image acquisition necessary to achieve the QIBA Claim. QIBA Profiles are based on published data whenever such data are available and on expert consensus opinion where no data exist.

Although based primarily in the USA, there are QIBA participants from North and South America, Europe and Asia. At the 2015 European Congress on Radiology, the European Society of Radiology (ESR) announced the formation of the European Imaging Biomarkers Alliance (EIBALL). In addition, leaders of the Japan Radiological Society (JRS) have met with the QIBA leaders to discuss future collaborations. Dr. Sullivan's lecture at the Fall Meeting of the JRS in October 2015 will provide more details about QIBA activities.

Key words

Quantitative Imaging Biomarkers Alliance (QIBA), Radiological Society of North America (RSNA), standardization, reproducibility, precision medicine

Rinsho Hyoka (*Clinical Evaluation*). 2017 ; 44 : W1-W22.

*1 This is the record of lecture at the 51th Autumn Clinical Meeting of Japan Radiological Society, held October 2-4, 2015, Morioka, Iwate, Japan. This lecture was provided on October 3.

*2 Liaison for External Relations, Quantitative Imaging Biomarkers Alliance (QIBA), RSNA.

Invited lecture on October 3, 2015, at the 51th Autumn Clinical Meeting of Japan Radiological Society, held October 2-4, 2015, Iwate, Japan.

Daniel C. Sullivan, M.D.

Dr. Sullivan is Professor Emeritus, Department of Radiology at Duke University Medical Center. He completed radiology residency and nuclear medicine fellowship in 1977 at Yale-New Haven Hospital. From 1977 to 1997 Dr. Sullivan was in academic radiology, holding faculty appointments at Yale University Medical Center, Duke University Medical Center, and University of Pennsylvania Medical Center, before joining the National Cancer Institute at NIH in 1997. From 1997 to 2007 Dr. Sullivan was Associate Director in the Division of Cancer Treatment and Diagnosis of the National Cancer Institute (NCI), and Head of the Cancer Imaging Program (CIP) at NCI. His areas of clinical and research expertise are in nuclear medicine and oncologic imaging, in particular focusing on improving the use of imaging as a biomarker in clinical trials. From 2007 to 2015 Dr. Sullivan was Science Adviser to the Radiological Society of North America (RSNA). During this time he founded and chaired the Quantitative Imaging Biomarkers Alliance (QIBA), which coordinates a wide range of national and international activities related to the evaluation and validation of quantitative imaging biomarkers for clinical research and practice.

Revised from the source web-site: http://nbdabiomarkers.org/bio/daniel-c-sullivan-md

1. Introduction

I would like to thank the organizers and leaders of the Japanese Radiological Society for this invitation. I'm very happy to tell you a little bit more about QIBA (Quantitative Imaging Biomarkers Alliance). Professor Tomio Inoue-has given you a very good background, and I will tell you a little bit more, so I would start with some of background because some people have asked me – What is the motivation for QIBA? What is this RSNA (Radiological Society of North America)?

I have been concerned about the issue of variability in radiology ever since I started in radiology about 40 years ago. I was concerned that when patients came to a hospital on one day they would get a particular answer depending on the machine that was used and the radiologist that was interpreting that study, and if they would have come the day before or the day after, they would get a different answer because there would be a different radiologist and a different scanner. In my academic work I began to do some studies about this variability.

2. Problem of variation and wrong scan interpretation

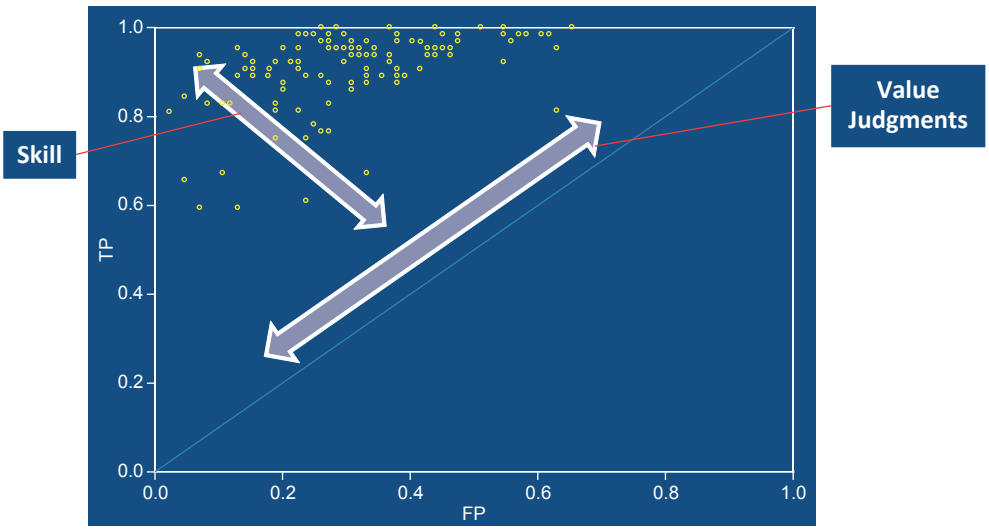
Variation and wrong scan interpretation have been a long-standing concern of mine. For example in the 1980s I published on the topic of "Effect of Different Observers and Different Criteria on Lung Scan Interpretation"*3. Years later when I was doing breast imaging and mammography I was still concerned about the wide range of variability and interpretation of the same mammograms.

So along with a couple of other colleagues, Dr.

Craig Beam and Dr. Peter Layde, we took a set of about a hundred mammograms and sent them to 108 different radiologists randomly selected around the United States, and had each of them read the same mammograms (Fig. 1). Their results in terms of sensitivity and specificity are plotted here on what's called an ROC (Receiver Operating Characteristic) curve. Each one of these points is one radiologist in terms of their sensitivity (Y-axis) and specificity (X-axis), and you can see they are all different. They're spread out all over the place. This phenomenon of radiologist variability has been studied by many people and has been documented in many areas of radiology. It's not unique to mammography or nuclear medicine. It occurs everywhere.

In my understanding of this problem over these many years, I came to understand that this variability depends on two main factors. So if you are one of these dots here, if you are one of these radiologists, where you're located on this graph depends on primarily two things. One of them is related to how good your expertise is compared to the random choice. So the diagonal line (Fig. 1) represents what the ROC curve would be if you just flip a coin every time you read a mammogram or some other study and made a decision based on the random flip of the coin. If you are an expert, perfect, you got every case right a hundred percent at a time – perfect sensitivity, perfect specificity – your point would be at the upper left-hand corner. So where you are between random and perfection depends on your skill and skill is related to both a combination of your congenital abilities related to vision and interpretation but also training and experience. That's where you are on the Y-axis. Where you are on the X-axis depends on your value judgment, your philosophy – whether you're a conservative or liberal in making judgments; whether you try to emphasize over-calling or under-calling. Both of these components of a radiologist's performance are very difficult to change. So if we want to try to reduce this variability just from the radiologist's

Fig. 1 Operating points of 108 radiologists reading same 100 mammograms



Beam CA, Layde PM, Sullivan DC. *Arch Intern Med.* 1996；156：209-13.

*3 Sullivan DC, Coleman RE, Mills SR, Ravin CE, Hedlund LW. Lung scan interpretation: Effect of different observers and different criteria. *Radiology*. 1983 Dec; 149(3): 803-7.

point of view, it's extremely difficult to change even one of these components. It's not impossible but very difficult. An easier approach really is to reduce the variability from the technical aspects of our image acquisition and extraction of information even though it is also difficult, and that is the emphasis in QIBA.

2.1 Premise of the problem

With that as a background, the fundamental issue behind QIBA is that variation in clinical practice results in poor outcomes and higher costs. And one way to reduce variability is to extract objective, quantitative data from the scans. I emphasize that this is only one way. There are many ways we can reduce variability and improve our standardization. But the one approach that QIBA focuses on is to improve the objective quantitative data that we get from scans.

2.2 Computers cannot replace radiologists

I want to say as an aside here that this does not mean that we think computers will replace radiologists. Certainly not in the near term, not in the lifetime of anybody in this room, because there are abilities that the brain and the eye system have in terms of visual recognition and judgment that computers cannot yet duplicate. Even though artificial intelligence is relatively advanced by some people's interpretation, it really cannot duplicate much of what the human brain does.

One example of that is the child's book that in the United States we call Where's Waldo. This is a task where the artist has hidden Waldo between lots of massive distracting images, and in this image, Waldo is here right in the middle. Waldo has a red and white striped shirt, red and white hat, brown hair, and blue trousers. The artist typically puts in other distractors like this one down here in the front of this picture. But even a 4- or 5-year old child can look at this and say immediately, "That can't be Waldo because that's a girl!" How the brain comes to that decision very quickly is not really known. It's not well understood and computers cannot do this. So it's a frequent exercise in computer software classes in graduate school to assign students to write an algorithm to find Waldo and typically the software cannot do it reliably. It's very difficult to do. This kind of extraction and judgment will continue to be important in radiology for a very long time. But that does not mean you can't improve image interpretation by using objective quantitative information in addition.

3. Motivation for QIBA

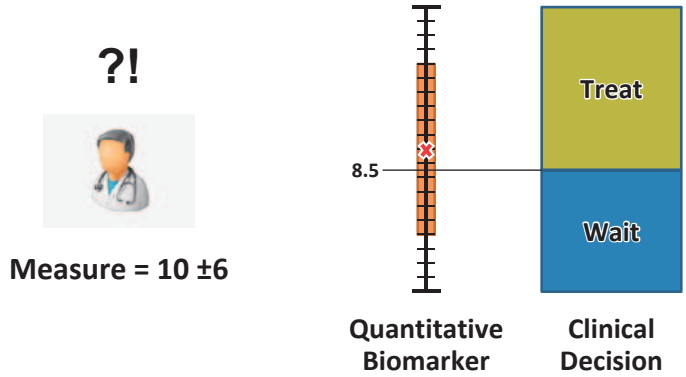
3.1 Quantitative imaging in healthcare

Quantitative imaging is not new. It's not something that has just been developed in the last few years. It's been around for 30 or 40 years or more in different forms – in ultrasound, vascular, cardiac (cardiologists have used numbers in their imaging studies for 30 years), in cancer, orthopaedics, and so on and so forth. So the concept is not new. But the issue which is a problem is that the measurements in radiology are not very precise (reproducible) and may not be very accurate.

3.2 Physician's use of quantitative information

This is a generic description of how physicians use quantitative information from any laboratory study, whether it's a blood test or potentially from imaging (Fig. 2). For example, let's say there's a blood test where the cut-off for normal is 8.5. If the blood test comes back lower than 8.5, the physician will do nothing. If the laboratory result is greater than 8.5, then the physician may do something. He will make some decision. He may order

Fig. 2 Motivation for QIBA



another test or some kind of a treatment. So in this case, if the measurement let's say is 10 (just hypothetically, see the red ✕ in Fig. 2), it looks like this is an abnormal result and that a treatment decision should be made. But in any test there is uncertainty of the number. If the uncertainty is large (in this case, let's say it's plus or minus 6) then the true value could be anywhere in this range from 4 to 16 (orange bar in Fig. 2). So it's not clear whether a value of 10 is really a normal test or an abnormal test. Hence, a physician is uncertain what to do. This is the problem with many of our measurements from radiology. There is a large degree of uncertainty and it is not suitable to make treatment decisions based on those numbers. So we need to improve that.

3.3 Variability in imaging measurements

I've mentioned that variability is due to many things and we can group them together as (1) the image acquisition variability in the scan, (2) the measurement method variability (whether one is using software or calipers) and then (3) variability of the radiologist. In QIBA, we take into consideration all of these because they all contribute to variability. But some are easier to deal with than others. I've mentioned that trying to reduce radiologist's variability is extremely difficult. So we focus on the image acquisition and measurement methods if we are trying to reduce variability.

3.4 Variability in scanner measurements

Scanners that are manufactured today are primarily manufactured to provide good images because that's what the radiologists want; that's what customers who are buying scanners want. The scanners are not necessarily good measuring devices because they're not engineered that way because customers don't demand that. They could be but right now they're not reliable measuring instruments. By reliable I mean the repeatability and reproducibility of measurement. It means that if we make the same measurement twice on the same patient with the same scanner, we should get measurement results that are very close to each other. They won't be exactly the same because there's always some variability, but they should be relatively close. That's repeatability – when the conditions of measurement are the same. There also needs to be reproducibility (comparability) which means if you take a different scanner – same patient but a different scanner – you should get a measurement which is very close to the measurements of the previous scan. In other words, measurements from different scanners and hardware should be very comparable. And for all these, it's important

to know how the whole system performs in terms of bias, precision, and linearity.

3.5 Poor reproducibility – Clinical implications

In the last few years, as people have paid more attention to extracting numbers from scans, there have been some studies that show that poor reproducibility of these numbers can have clinical implications; can have effect on patients. This is the issue that has troubled me for 30 or 40 years. For example, one paper from *Radiology* in 2014 looked at coronary artery calcification scoring with different scanners from different vendors*4. The authors concluded that the results from scanners made by different vendors produce substantially different scores which can result in reclassification of patients to high or low risk categories in up to 6.5 percent of cases. Again, in other words, if a patient came in one day and had a scan from manufacturer A but happens to choose to come to the hospital the next day and got the scan from a different vendor, they could get a different score and the result for that patient could change from high or low risk category just depending on what day they came to radiology department and what scanner they got.

This is a similar study showing the reproducibility of non-calcified coronary artery plaque quantification from CT (computed tomography) using different image analysis software*5. In this case, the variable that changed was the software. The conclusion though is similar, i.e. currently available quantification software provides poor inter-platform reproducibility. Serial or comparative measurements would require evaluation using the same software and industry standards need to be developed. That's what they've concluded. These two examples show one example of problems with hardware and one example of problems with software. That's all part of the issue we have to deal with.

4. Quantification – Consumer expectations

In the last few years, it's become apparent that treating physicians, referring physicians, increasingly want to use quantitative results. They want quantitative objective information partly because the amount of information that's available to physicians of today is enormous, and in order to integrate all that information, to have it in objective quantitative form is very helpful. One of the areas that is being developed increasingly is called *decision support algorithms* that can integrate information from a variety of sources. Typically however, the information in radiology reports nowadays does not lend itself to be included in decision support algorithms because it's in narrative text, and computer algorithms cannot deal with that well at all. Radiology reports right now are the worst part of the electronic medical record in terms of being standardized, objective and quantitative.

4.1 Publications show need for more quantitative information

There are many publications from referring physicians or other users of radiology indicating that they want more quantitative information. For

*4 Willemink MJ, et.al. Coronary artery calcification scoring with state-of-the-art CT scanners from different vendors has substantial effect on risk classification. *Radiology*. 2014; 273(3): 695-702.

*5 Oberoi S, et al. Reproducibility of noncalcified coronary artery plaque burden quantification from coronary CT angiography across different image analysis platforms. *AJR Am J Roentgenol*. 2014; 202(1): W43-9. doi: 10.2214/AJR.13.11225.

example, there is a paper from *the American Journal of Radiology* indicating that 94% of oncologists expect measurements of tumors*6. In 2010, the American Thoracic Society and the European Thoracic Society jointly issued a policy statement that says in effect – We know that CT scans can provide quantitative information and we want that quantitative information to help us make treatment decisions in emphysema*7. One of the thresholds that the pulmonologists have said they think they need is that they would like to be able to see the difference of all the lung density changing by one or two percent – that that would be clinically meaningful in terms of therapy for patients or for any clinical trial to potentially determine whether the therapy was useful. I mention that threshold because I'm going to come back to that. It's an example of a difficult threshold. Hepatologists (liver experts), medical doctors that focus on liver disease want quantitative measures of liver fat infiltration*8. Rheumatologists want quantitative measures on joint disease*9. Neurologists and psychiatrists want a variety of quantitative measures of brain disorders because without objective measures they have to rely on subjective symptoms that the patient describes which are very variable. In the United States, our regulatory agencies increasingly want more objectivity in imaging scan interpretation for approval for marketing or approval for reimbursement.

4.2 Research on quantitative imaging is increasing

Also, it's obvious that a lot of research about quantitative imaging has been going on in the last few years. I did a recent search in PubMed on terms related to quantitative imaging biomarkers and there are thousands and thousands of research publications about quantitative imaging. Moreover, when I look at any imaging journal table of contents, it's apparent to me many of the articles are related to quantitative imaging month after month. So to quantitate that, I looked at the journals from RSNA in 2014 (*Radiology* and *Radiographics*) and I found that of the total 750 articles, 250 or one-third, have the word quantification in the title. So people are continuing to publish a lot about quantification. Of course, many of these articles actually come from investigators in Japan. I recently looked at one of your journals, *Japanese Journal of Radiology* which I understand changed to English about 10 years ago, and just a quick search of the word quantitative imaging in the English language journal showed about 195 articles. So just in this one Japanese journal there were that many articles written on quantitative imaging in the past few years. And of course, as I said, many Japanese investigators have published elsewhere. Hence, it's obvious that a lot of this work is going on in Japan as well as elsewhere in the world: in the United States, Europe, Australia, South America, people are working on this.

---

*6 Jaffe TA, Wickersham NW, Sullivan DC. Quantitative imaging in oncology patients: Part 2, oncologists' opinions and expectations at major U.S. cancer centers. *AJR Am J Roentgenol*. 2010 Jul; 195(1): W19-30.

*7 Hsia CC, Hyde DM, Ochs M, Weibel ER; ATS/ERS Joint Task Force on Quantitative Assessment of Lung Structure. An official research policy statement of the American Thoracic Society/European Respiratory Society: Standards for quantitative assessment of lung structure. *Am J Respir Crit Care Med*. 2010 Feb 15; 181(4): 394-418. doi: 10.1164/rccm.200809-1522ST.

*8 Fitzpatrick E, Dhawan A. Noninvasive biomarkers in non-alcoholic fatty liver disease: Current status and a glimpse of the future [review]. *World J Gastroenterol*. 2014 Aug 21; 20(31): 10851-63. doi: 10.3748/wjg.v20.i31.10851.

*9 Chu CR, Millis MB, Olson SA. Osteoarthritis: From palliation to prevention: AOA critical issues [review]. *J Bone Joint Surg Am*. 2014 Aug 6; 96(15): e130.

4.3 Current commercially available MR QIB applications

Some of these research studies have turned into commercially available applications. There are many CT, PET and MRI quantitative imaging algorithms commercially available. Some of them are probably on display here at this meeting.

5. Impediments in using quantitative imaging

Yet few people in radiology are using quantitative imaging. There's a lot of research and publications about quantitative imaging. There are a lot of commercially available products, and yet radiologists are not using it in general. They're not using it very much. So why is that? What's the reason? What's the problem? There are several things that contribute to this and some of them have to do with precision and reliability but the main one overall, in my judgment, is that there are very few clinical treatment decisions currently driven by quantitative imaging results. Physicians don't generally make a clear-cut treatment or clinical decision based on whether the volume of a mass on a CT or MRI is at a certain threshold or whether the SUV (standardized uptake value) of a PET scan is at a certain threshold. To some extent those numbers are being used but there is hesitancy about making treatment decisions because there's not yet good data about the reliability of making those measurements and the reliability of making treatment decisions based on those numbers. Also, the lack of industry standardization prevents us from obtaining such clinical evidence. So this is like the idiom in English we call the "chicken-and-egg situation" – two things that are dependent on each other; difficult for either one to go forward because the other one has to occur first.

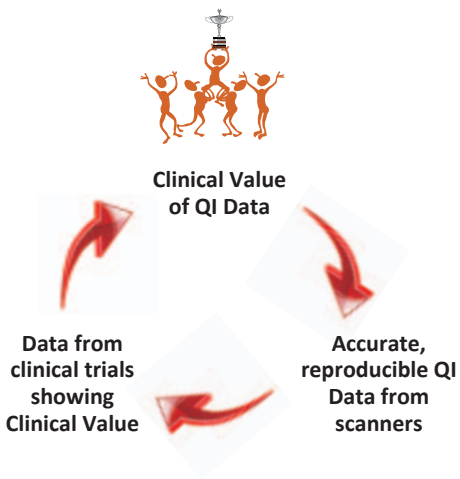
5.1 Lung densitometry example

Earlier I mentioned the lung densitometry as an example. That pulmonologists said that they would like to be able to reliably know if there is a change of one percent in overall lung density. But there's no therapy for emphysema now; so they would not be making any decision for the patient based on that kind of a change if we could produce it. And it would be very difficult right now for CT scanners to have this kind of reproducibility of one percent in overall average of Hounsfield units. But if we could get precision to a level of 2 to 5 percent in a clinical trial where these numbers would be used to separate groups – not to make an individual patient decision but to separate groups into responders or non-responders – that would be very useful; and that would be a first step towards establishing the clinical utility of this kind of a number.

5.2 Toward quantitative imaging

We are going about this in a stepwise or what we call an iterative fashion. Our overall goal is to establish the clinical value of quantitative imaging data. But to do that we need good, reliable data from clinical trials to show the clinical value; and

Fig. 3　Toward quantitative imaging



in order to get good data from clinical trials, we need to have scanners that provide accurate and reproducible numbers of a known precision and known reliability. When we talked to the manufacturers about that they say, "Well, our customers don't ask for that because it's not clear that there's clinical value in those numbers." So we have these interlocking problems that cannot be solved in a linear fashion one after the other. They have to be all attacked simultaneously in a way. Our approach in QIBA is to try and attack these problems from multiple points at a time and then gradually improve and work towards an ultimate solution which is much better than our current situation (Fig. 3).

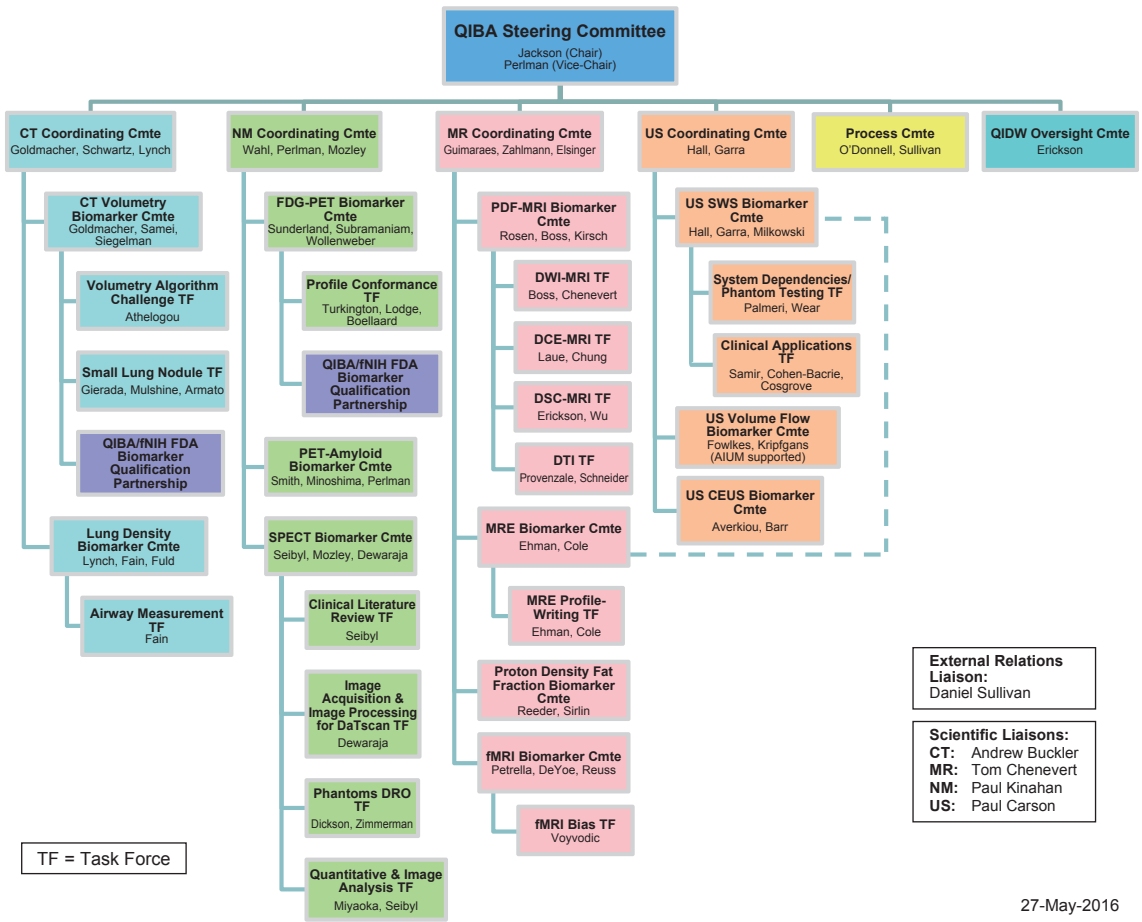
6. Quantitative Imaging Biomarkers Alliance (QIBA)

Just to repeat the background, we started the RSNA in 2007. Our overall goal is to get the manufacturers to build imaging devices that are also measuring devices. There's lots of information here on this website (http://rsna.org/QIBA.aspx).

6.1 QIBA organization chart

Our organizational chart can be found here (Fig. 4 ; http://www.rsna.org/uploadedFiles/RSNA/Content/Science_and_Education/QIBA/QIBA-

Fig. 4 QIBA organization chart



Organizational-Chart(1).pdf). Our biomarker committees are grouped by modality – CT committees, MR committees, nuclear medicine (NM) committees, ultrasound committees. Each biomarker, which is a particular measurement that we're interested in, has a committee that focuses on that. There are many biomarkers that are not yet on this chart so there's plenty of opportunity for more people to get involved and develop committees to focus on other biomarkers. Also, under each committee there are 2 or 3 co-chair names that are listed. The reason for this is that we try to have one co-chair who is a radiologist, one co-chair who is a physicist or an engineer, and one co-chair from industry. We don't have 3 for every committee but as much as possible we try to have co-chairs that represent different stakeholder groups because it is very important to always have different perspectives involved, and the perspectives of physicists and engineers are absolutely critical and involvement of industry is absolutely essential. All of our committees are open to anybody who is interested. They meet by conference call usually every couple of weeks, and anybody can join on those conference calls. You can go on the email list and receive information about those calls. I realize that because of the time difference between the United States and Japan it would be very inconvenient for most of you to join in these calls but in some cases it might be possible.

6.2 QIBA meeting at RSNA annual meeting

I also want to mention that we have a QIBA session every year at the RSNA annual meeting. So if any of you are planning to go to the RSNA annual meeting this year there is a general plenary session which will include a QIBA overview and updates and then there will be a breakout session for the individual committees. For example, you could go to the CT volumetry committee or the MR biomarker committee or whatever committee you might be interested in or you could move from one committee to another after an hour to hear the different perspectives. They are open to everyone. If you want more information about this, such as room number and so forth, you could email me or email RSNA staff on their website address and we would be very happy to give you the information.

6.3 QIBA approach

You've all heard that the approach in QIBA is to identify sources of variations for particular measurements, to specify solution which is in the form of a document that we call "Profile", to test those solutions, and to promulgate or disseminate those profiles or solutions to users and vendors.

6.4 QIBA profiles

QIBA profile is a long document comprised of many pages. It's a systems engineering document that describes everything that has to be done to achieve a claim.

In PET SUV for example, these are all the steps (Fig. 5). Patient preparation, scan acquisition, image reconstruction, image analysis and interpretation all have to be standardized in order for the SUV to have a certain known precision and be reproducible. If you're going to look at the change in the biomarker (which is usually what we're interested in in radiology), let's say a change in SUV after therapy, then all of these things should be standardized at both time points for both scans in order for the change to have a known reproducibility (precision).

There are many stakeholder groups and it is essential to have them involved.

6.5 Bias and precision

A few years ago after QIBA started, it was

Fig. 5 FDG-PET SUV example

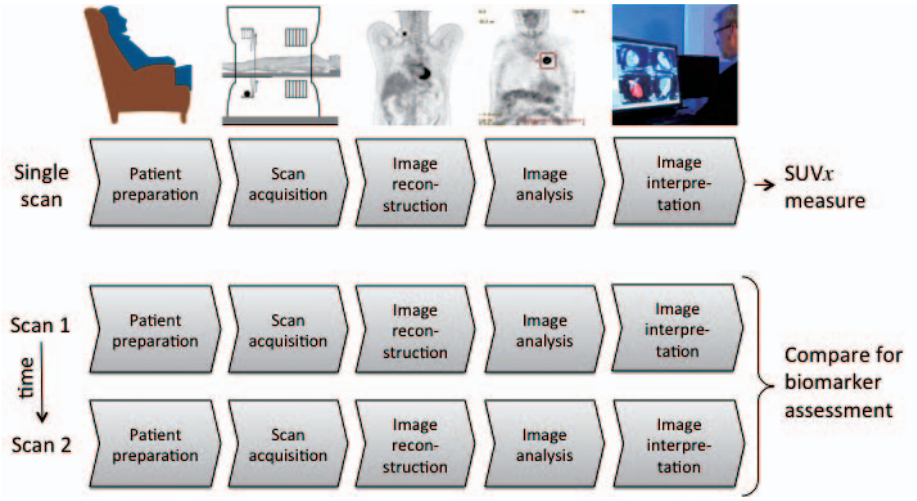
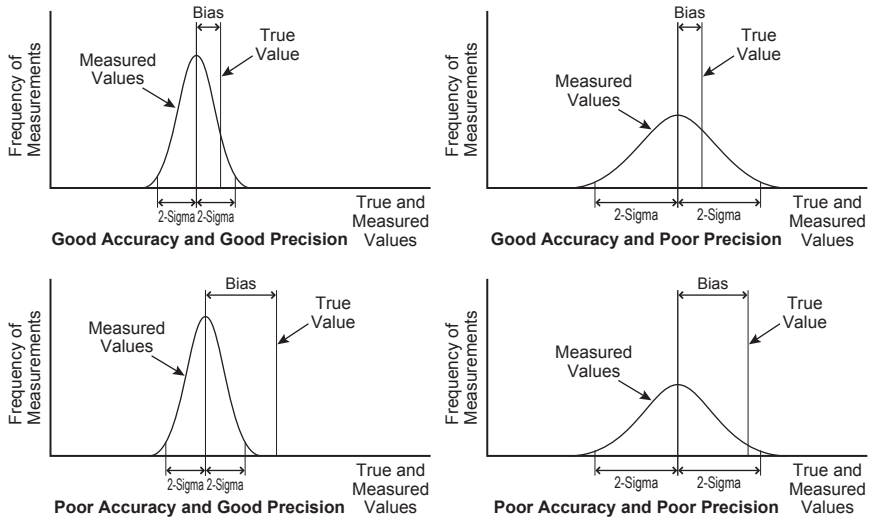


Fig. 6 Bias and precision



Bias may cancel out in serial measurements.

apparent that people who are involved in discussion were not using statistical terminology in a consistent way and sometimes it was ambiguous. So we convened a group of measurement experts/methodologists to advise us about the statistical measurement terms that we should use; what the definitions should be and what the minimum parameters should be for us to say that we have characterized the biomarker. Their recommendation was that we should focus on bias and precision (Fig. 6). These are the two parameters that we need to understand at a minimum to understand the biomarker, to standardize it. In English, the measurement scientists said that they prefer the word "bias" instead of accuracy. Some people use the terms for accuracy and bias interchangeably, but the measurement sci-

entists have said accuracy is not as clear as bias. So these are the preferred terms and I just want to review the definitions with you, for those of you who don't use these terms every day.

"Bias" is how far the mean of multiple measurements is away from the true value, the actual value of the measurement, and "precision" is the spread of multiple measurements of that same entity. So bias could be small which is good or it could be large which is not so good, and precision could be small which is good or large which is not so good. There can be any combination of these four possibilities.

Now in clinical medicine, actually in any measurement, in order to know what the bias is you have to know what the true value is, and in clinical medicine most of the time we don't know what the true value is. We cannot really measure the true value. We measure the true value in a test object or phantom but we cannot measure it in a patient. Often you don't know what bias is in a clinical setting. We can assume what it is from phantom measurements but if we are measuring two time points in the same patient and if we know that bias is linear, then bias cancels out and we don't have to worry about it. But you have to worry about the fact that we don't know what it is. And so bias may cancel out. Hence, the most important parameter in biomarkers is "precision". What's the test/re-test reproducibility; that's the variability you get when you make the same measurement from the same patient twice; and that could be with using the same scanner, the same software or not.

6.6 QIBA profile structure

The structure of a QIBA profile can be found here (http://www.rsna.org/QIBA-Profiles-and-Protocols/). I don't think we need to go into that in detail but I want to spend a couple of minutes talking about what we call the "claim" that the profile is based on.

6.6.1 Examples of QIBA claim statements

The recommendation from our measurement experts was that a claim would say – there is a 95 percent probability that the measured biomarker plus or minus the precision encompasses the true value. An example in PET would be there's a 95 percent probability that the measured SUV plus or minus 15 percent encompasses the true SUV value. I mentioned that most of the time we're interested in change in imaging; so the claim that's related to change is very similar. But it's really like this – there is a 95 percent probability that the measured change in the biomarker plus or minus the precision encompasses the true change. An example might be – there is a 95 percent probability that the measured change in mass value plus or minus 30 percent encompasses the true change in mass volume. Some of you may have trouble understanding the nuances of this statement, so I want to explain this again in a couple of different ways.

I'll start with the SUV as an example. Let's assume we have a patient with cancer (like lung cancer) who has a PET scan, and the SUV of the lung cancer measured 6. The precision was plus or minus 15 percent. That means the true value, true SUV, is really somewhere between 5.1 and 6.9 (15 percent of 6.0 is 0.9). We don't know what the true value is. We have 95 percent confidence that it's somewhere in the range of 5.1 and 6.9. But we don't know exactly what it is. Now, let's presume the patient comes back a month or two later after therapy, and the SUV measurement is now 5. It seems to be getting better. However, because of precision (variability in the measurement), the true value is somewhere between 4.25 and 5.75. So this range – 5.1 and 6.9 – overlaps with this range – 4.25 and 5.75. It's possible for example, that the true value here before therapy was 5.5 and the true value here after therapy was 5.5. Nothing has really

changed biologically even though the measured SUV appears to have changed, and at first glance you might be inclined to say, "This is getting better." But we really cannot say that. We don't know that. As a general rule of thumb, there has to be a change of twice the precision in order to feel confident that there's been a biological change. That's why the precision numbers are important because as a rule of thumb, on average you could say, if the measurement has changed by at least a factor of two times the precision then we can have confidence that there is a biological change. For example, if this patient came back with an SUV of plus or minus 4.2, now the true value would be between 3.6 and 4.8. This is different than the range from the original (pre-treatment) value, and so we can say with more confidence (95 percent confidence) that this patient has a biological change; that this change from 6 to 4.2 is not just due to the variation in the measurement of the scanner and the software but it is a true biological change. This is the importance of precision and we have to work it out pretty carefully. Additionally, clinical trials have to be done to determine whether this biological change is clinically important and can be used to make clinical decisions. That's a whole additional set of studies that is beyond the scope of QIBA.

Now in that example, I used the precision of plus or minus 15 percent for all of the measurements. Unfortunately, life is not that simple and precision does change depending on other things that are or may have changed at the time of the measurement. The precision will be different if we change scanners or if the scanners are not manufactured to have exactly the same precision. Software could be a factor. The size of the lesion could be a factor. In the CT volumetry, for instance, precision will vary depending on the size of the nodule, and all those other things – scanners, software, and reader – used in the same time points.

6.6.2 Expected precision for alternate scenarios

In the QIBA profile for the CT volumetry, we have more information about the precision and how it changes, and we developed a table that gets quite complicated. So for example, if you make two measurements on lung cancer on CT scan – or a nodule if you're doing CT screening – and you use the same scanner at both time points and the same radiologist reads them using the same software, then the precision is plus or minus 11%. But if you use different scanners and different radiologists make the measurements and they use different software, precision really becomes very poor. It's only plus or minus 47%.

Very often in a clinical setting a patient may have the two scans. The two CT scans might be done in two different scanners (because it may be difficult to get them on the same scanner) but the same radiologist has access to both studies. So even if the first scan was measured by a different radiologist, then the radiologist doing the reading now can redo the measurement using the same software. So the situation really is different scanner but same radiologist, same software, and so the precision now is plus or minus 32% or about 30%. That's why on the previous slide I used the example of plus or minus 30% for the CT example because in a clinical setting that's likely to be the example. However, this table only takes into consideration whether the scanner, software and radiologists are different.

6.6.3 CT volumetry example – coefficients of variation

There are unfortunately more variables, for example the precision changes depending on the size of the nodule. Precision is worse for smaller nodules and better for bigger nodules. So if a nodule is increasing in size, precision of those two measurements is going to change and it may also

change depending on where in the bore of the CT scanner the nodule is located. Precision will be better at the centre and worse at the periphery. Let's say during the baseline CT scan the nodule is 8 millimeters, and if the nodule measures 9 millimeters when the patient comes back for follow-up scan the variance associated with those two measurements overlap. The difference between these two measurements of 8 and 9 is not two times the precision. So you cannot say with certainty that the change from 8 to 9 represents a biological change. If, however, the patient measures 10 millimeters or more on follow-up scan, then that is a biological change from 8 with 95% confidence.

But again this has to take into consideration a lot of these factors which are too complicated for any radiologist to bother with frankly. It's too much to deal with in clinical practice, so some of the people interested in this have formed a small company of software experts and developed a "calculator" which is like a spreadsheet that has some formulas behind it, and this is available on their website. You can go and look at this from this website – http://accumetra.com/NoduleCalculator.html. This allows the radiologist to take the measurement at the first time point in the volume and then take the measurement at the second time point and some of the other factors that we talked about are incorporated into this calculator, and it calculates the difference for you and then will tell you whether this really represents a true change. It will tell you whether or not this is a true change with 95% of confidence. As we learn more in QIBA about these issues, we're learning that precision is more complicated in some cases than we thought. There are similar issues in ultrasound, MRI, PET, and there's a lot of work that still needs to be done to develop the data that we need to work out these kinds of calculators.

6.6.4 QIBA profile: Actors, Activities, Requirements

The Profile describes all of the people and software that are involved as "Actors", which is the word that we use in English to refer to all these devices or people that are referred to in the Profile. What they do or what they have to do, are called "Activities" and those are all listed in the profile and then under each of these there would be the specifications or guidelines of what has to be done.

6.6.5 Conformance to QIBA profile

Anybody who is using or wants to use these measurements will want to know – Does this vendor, does this scanner, conform to a QIBA profile? Does the site conform to a QIBA profile? So one of the activities we have in QIBA is to develop methods for use by vendors to certify that they are in conformance, and it could be at one of three levels. One would be that the vendor does the test himself and says my device conforms. There's no external or objective confirmation of that fact. The second level would be what we call community corroboration which means, for example, a physicist at some site could do the test again and say – "Yes, I agree. I corroborate that this scanner does conform." The third and the most formal would be a formal certification process by a third party entity like the American College of Radiology or RSNA or some other entity that right now doesn't exist because it would require more infrastructure, money and so forth. But that's one of the things that we are exploring.

7. MR variables

Now I wanted to just change topics here for a minute and talk about some of the help that we have for working on these things.

We are fortunate to have collaboration with the

National Institutes of Standards and Technology in the United States which really is a group of physics and engineering experts. They have two sites in the United States. One is in Boulder, Colorado and the other one is near Washington, DC in Gaithersburg, Maryland. The MR group is located in Colorado. The CT and PET group is located in Gaithersburg. NIST is very interested in helping us and very active in helping us in developing the test objects, phantoms and other standards for use in the Profiles.

8. QIBA funds

We have some funds from the NIH/NIBIB, National Institute of Biomedical Imaging and Bioengineering, to help us fund some of these projects. We get from them about a million dollars a year and we're able to fund about 15 projects each year in a range of about 50,000 dollars per project to work on developing test objects or doing a data collection test.

9. Dynamic contrast-enhanced MRI (DCE MRI)

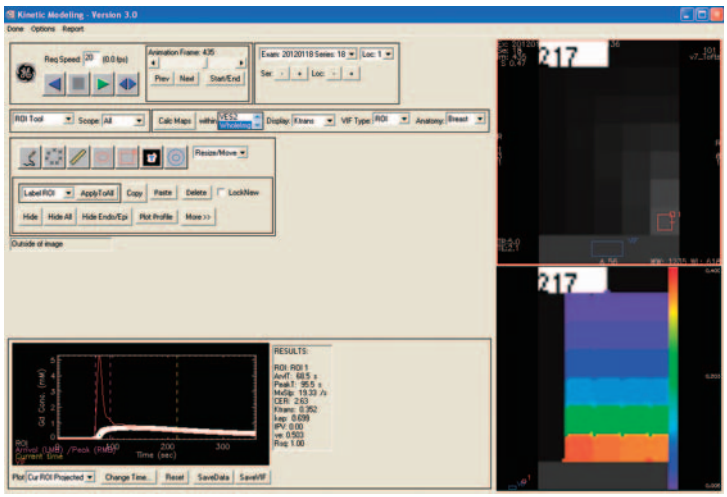
For example, we have developed a DCE (dynamic contrast-enhanced) MRI phantom. An important point of this is that in addition to the phantom we have an analysis tool. We have an image archive, image warehouse website, where the images can be uploaded to the website. Then the person who is doing the image can run the software against their image, and it will then automatically detect the various locations in the phantom and calculate whether they are consistent with the standards or not.

In addition to physical test objects, several QIBA committees create synthetic data which then can be processed by a work station or software to test the work station and software. One example is the synthetic data for DCE-MRI (Fig. 7). This is freely

Fig. 7 RSNA QIBA DCE-MRI Digital Reference Object (DRO)

(Barboriak)

- **Simulated T1 measurement data for range of S_0 levels and added noise levels**
- **Simulated DCE measurement data for range of S_0 levels**
- **Simulated DCE measurement data for extended Tofts model**



available from our website. Anybody can download it. People all over the world can download it to the software and run it to test their algorithms or to test their work station.

We also have automated software, so after you run this synthetic data you can upload the results to our QIBA image warehouse and run this automatic software package. The synthetic data was created by one of my colleagues at Duke (Daniel Barboriak) but the automated software package was created by one of our QIBA colleagues (Hendrik Laue) in Cologne, Germany. This is an example of an international collaboration that we have in QIBA committees. So we're very interested in having more of this type of international collaborations.

10. Biomarker development

As I mentioned, QIBA focuses on the technical performance of the assay. The clinical performance is beyond our scope because we don't have much money. This is another opportunity for other groups, for example in Japan, maybe clinical trial groups, to get interested in testing the QIBA profiles or collecting precision data.

11. European Imaging Biomarkers Alliance (EIBALL)

In Europe, the European Society of Radiology in the past year has organized some activities into what they call the European Imaging Biomarkers Alliance (EIBALL) which is very similar to QIBA. It's the European version of QIBA. EIBALL have made a decision to try to focus on MRI because that's their area of interest right now. But they will maintain close collaboration with QIBA to avoid duplication and benefit complementarity. One of the ways we do that is appoint one of the chairs of EIBALL as member of the QIBA steering committee and QIBA has also appointed QIBA chair (now Ed Jackson) as the QIBA representative to EIBALL. These are just a few examples as you think about opportunities for some QIBA activities in Japan. There are lots of possibilities. It doesn't have to be this kind of model. It can be a new model but these are some examples of how it could be done.

12. QIBA-related activity in Japan

In Japan, a committee was established in Japan Radiological Society (JRS) to participate in discussion of the QIBA, to which some of the member radiologists were assigned. This committee was established by Professor Tomio Inoue's enthusiastic initiative and they are also engaged in how it is possible to introduce standardization in research and practice in Japanese situation.

13. Importance of obtaining clinical precision data

Lastly, I wanted to mention that one of the areas that is of critical importance is to get data about precision and I tried to emphasize that in the talk. It is important to note that it only takes a small trial in the range of between 15 to 30 subjects to provide useful precision data. Very often though when people think about a clinical trial to show whether imaging is useful or not, they think it may have to be a hundred or several hundreds of patients. But for test/re-test of precision data, it doesn't have to be that big – fifteen to 30 subjects is useful. One of the ways that that can be useful is to combine several studies by using meta-analysis if the methodology is standardized and all the information is published. These could be independent studies that you can do on your own but it could be sub-studies nested in larger clinical trials.

14. Conclusion

I will conclude by saying some imaging biomarkers are ready for use and could be used more widely but many of them need more research and development to further reduce bias, improve precision and assess the clinical utility in clinical trials. Some of the things you can do to help is, first of all embrace standardization across the medical centers, collect precision data in clinical trials (it's critically important), and participate in QIBA. We would be very happy to have you work with us, and we'll be happy to offer you a lot of information and assistance.

＜Q&A＞

● Standardization and software

Q There are several vendors which provide CT human tissue volume measurement. Now we have commercially available software for this. Do you think it is easy to control or regulate these kinds of software to make it standard?

Sullivan Yes. That's a very good point. You used the word regulate. In the United States, the word "regulate" has legal implications, so we try to avoid using that word. What we try to focus on are standards and methods for a vendor to certify that they meet a standard. One of the ways they could do that would be to have a large archive of CT scans with masses and anybody could run software against this library of CT scans and see what the results are. There are some libraries of CT scans where that could be done. But up to this point, most of those libraries of CT scans are so freely available that the vendors have actually used them to create their outputs so it's not a good test. So one of the things that we're doing right now is we're using some of the money that we receive from NIH to fund some groups to create "synthetic" lesions by extracting real lesions from some of the CT scans and create a library of lesions which then can be superimposed on new CT scans. Those libraries of lesions can be changed in any way by the computer – make it smaller, change the shape – and they can be superimposed into the CT scans in any location. It's a way of creating continuously a new library of CT scans that the vendor could not have ever tested or could not have used to create a measurement. So it would be a more valid test. I think that that's going to be a very good way to test that specific issue. Today we have about 200 of those lesions that have been created into a library and we're now in the process of testing that concept. Right now it's just a concept. But I think maybe within the time course of a year that will be available on our QIBA website for manufacturers to test. That's a specific example about the CT volumetry that you've raised. But that's also a general issue, a general concern – How do we create testing environment for any algorithms? We think this is a useful concept and maybe it could be applied to other things. It's a general issue that we have to try to figure out how to do that.

Q We have at least 4 or 5 CT volumetry software, Japanese software. I think now we can use these softwares very easily.

Sullivan I would just add one more comment because you used the word regulate. In the United States, the FDA (Food and Drug Administration) is the group that does the regulating and they are very interested in this whole issue as well. They're trying to figure out how to deal with it. They're very interested in QIBA. They participate in many of our QIBA committees. Some people from the FDA are involved in that project that I've just described with the nodules, and they're involved in many of our other committees as well. So the FDA

is trying to figure out the best way to do this. Within the next couple of years, I think they will also have some public meetings with industry to talk about it; how to develop some standards for this.

● Commitment of industries

Q Industry is one of the most important players for the standardization of imaging modalities and record modification. However, industries are sometimes reluctant to standardize because they think that some of their own technologies are interfered with such unique platform of standardization. My question is what strategy do you use to involve industries in the framework of QIBA?

Sullivan Yes. Again you've raised a very important point, and this was an issue we recognized early on in 2007-2008 when we started QIBA. We knew this would be a problem, and we made many trips to all the companies to talk to them individually about what we wanted to do and to get their input. And you're exactly right; at first they were resistant, skeptical, and very concerned about their proprietary devices or software and their ability to differentiate themselves from each other. But things have changed over the last 6 to 7 years, because they recognized the medical importance of getting the same answer. They understand that they all should get the same answer; just like a thermometer or a blood pressure cuff or any other device. Hence, our general approach in QIBA is similar to what I've just described, that is we don't tell vendors how to get the answer. The QIBA Profile provides a means and a test for the vendor to show that they get the right answer. How they do it is up to them. They don't have to tell anybody how they do it. They may retain the privacy of their intellectual property. But they have to show that they can get the right result. That's one of the values, for example, of the consistency of synthetic data that I showed earlier. I showed the one for DCE-MRI which has been very useful. But we have another one for PET, for SUV/FDG. When that was made available about 3 years ago, many vendors ran that software through their workstations, and many of them found they got the wrong answer. They were surprised because there never was a way to test this in the past. So they have now fixed it. We don't know what the problem was. They don't have to tell us what the problem was, but they have fixed it and now they get the right answer. That's the general approach. We tell them what they have to get at the end; not how to get there. The vendors in general have increasingly been very supportive in all of these areas. For example in CT, although we talked about CT volumetry, CT densitometry has some different issues which I didn't realize until we got into QIBA. When I was a younger radiologist I did not realize that the Hounsfield unit does not have a single, unambiguous definition. It's a relative measurement depending on the quality of the beam. So to get CT densitometry the same across all scanners is actually very difficult because how the Hounsfield unit is calculated is different. So now we have a group of representatives, engineers, from all of the four vendors. They created this group on their own and they are making measurements from the same phantom to see what they get. The purpose of this is to see if they can develop a normalization factor that would apply to individual scanners to get the same result. They don't have to have the same x-ray beam spectrum. They could have whatever they want (e.g. spectrum). They can have their own reconstructions but when using the same phantom they should get a Hounsfield-unit measurement of air that is within the standardized range. This is another example of how industry has been very helpful in the last few years trying to solve these problems on their own.

Left: Daniel C. Sullivan; Right: Tomio Inoue

＜Additional Q&A session＞[*10]

● Importance of quantitative imaging for drug development

Q I would like to ask you how important this kind of quantitative imaging, quality assurance and standardization activity is, especially for drug development, and how important governmental funding is, and also consortium development with government and companies and academic institutes.

Sullivan The basic problem to deal with is that most of radiology image interpretation for many decades has been subjective. And in any kind of research setting, like drug development, that's not useful for establishing the result of an experiment. A drug trial is an experiment. And so you need to have objective data to determine what's the valid conclusion from the experiment. So to use imaging as an output or an end-point in a clinical trial, it's necessary to have some objective, quantitative result from imaging. And it's very possible, very feasible to do that now because all images are digital: every pixel, or every voxel in an image is made up of numbers.

But it's also necessary for those numbers to be useful; there has to be standardization of these images because there are many factors involved in creating the image. There are patient factors, scanner factors, software factors, image processing factors. So all of them have to be standardized in order to get the numbers that are reproducible. And there are a variety of quality assurance and quality control programs to try to help with that, and a variety of guidelines. QIBA is one of them, and the main purpose of QIBA is to really focus on the precision or reproducibility of imaging.

● In the era of big data and precision medicine

Q So how is the perspective to using imaging in

[*10] This additional interview was separated discussion after the lecture of Professor Sullivan, interviewed by Tomio Inoue and Chieko Kurihara.

the era of "big data" and "precision medicine", as it is different trend from the previous "critical path" initiative as a governmental funding? For example, we regard that in precision medicine data sharing strategy by means of imaging archive is very much important and in such situation standardization and reproducibility is coming to be more and more important.

Sullivan Yes, absolutely. For many years, people have been interested in biological specimens for a variety of purposes and especially genomics. There's been a lot of emphasis on developing biobanks – bio-repositories. The need for an image archive is exactly the same as a need for a biobank or a bio-repository. The need to collect hundreds of thousands of images in order to analyze and process all those images and determine what biomarkers are useful. In the past it was technically difficult because the scans, MRI scans or CT scans are very large in terms of their file size. They require huge amounts of storage and software to move huge files around. Now those technical problems have pretty much been overcome. So now really the limitation is partly money; there needs to be an ongoing source of money to keep supporting this for many years because you have to, just like a bio-repository that you keep in a refrigerator for years and years. And there needs to be some entity like a neutral, trusted organization to manage this. It could be the government or it could be a professional organization or a consortium. It's a problem in the United States. I think in Europe, the European Society of Radiology is trying to deal with that also; to create a large databank. I don't know what's happening in Japan but it's a problem to get an organization and the money. Technology is not a problem.

● Reproducibility and standardization in early and late phases

Q During clinical development process from early phase hypothesis generating until late phase confirmatory study aiming at product approval, there should be different critical points of standardization. Everybody can understand it is very much important to standardize, at the time of approval process, the later phase of clinical development. And also after the approval, these people are very much interested in the standardization in the clinical practice. Meanwhile, how do you think about the early phase where scientific purpose is hypothesis generation?

Sullivan I think that standardization is much more important at the end of the process, the Phase 3 trial, the definitive trial, and in maybe post-marketing approval or examination. In the early stage, standardization is much less important. What's more important at that early stage is to report carefully everything about whatever you did so that somebody else could reproduce it. So reproducibility would be important. So that you can have more opportunity for innovation and development. This usually requires publication of lots of information including maybe more than you can put in the article; so that there's a need for supplementary information.

● Visual radiologist

Q Additional question: what is the meaning of "visual radiologist". In QIBA, people say that radiologist should be visible. We often hear two keywords: One is "precision medicine", and the other word is "radiologist should be visible".

Sullivan In the United States there is an emphasis now on radiologists consulting with the

patient, giving the results to the patient, and consulting more with the referring physician. Because digital images can be transmitted from where they were acquired to radiologists in a distant location, and because the reporting systems are all digital, the radiologists often are physically very remote from the patients and the referring physicians. They may not see each other and this creates problems. So the American College of Radiology now is promoting or pushing the idea that radiologists should go talk to the patient; give the patient results, and have consultation and talk to the patient, and talk to the consultant.

Q Imaging diagnosis would be critical role in the diagnosis of patient rather than each specialists – cardiologist, cancer physician.

Sullivan In some places in the United States now a patient can schedule a consultation with the radiologists directly. They can just come in with their reports and say, you know, "explain this to me."

Q Previously radiologist is the physician's physician. But it's changed.

Sullivan Yes. Previously they never talked to the patient. Now it's changing.

● Relationship between UPICT and QIBA

Q Our understanding is that at the time of starting to develop UPICT (Uniform Protocols for Imaging in Clinical Trials), the Society of Nuclear Medicine (now Society of Nuclear Medicine and Molecular Imaging) played a key role, and this first meeting was held during the Annual meeting of SNM, held in 2010, San Antonio*11. Later on, UPICT development came to be an activity of QIBA. However you said that QIBA started around 2007 to 2008, and UPICT developed before QIBA. How is the relationship between QIBA and UPICT; SNM and RSNA?

Sullivan The UPICT really was started when I was at NIH in around maybe 2001 or 2002. We started UPICT because we realized in the cancer clinical trials things were not being done in a standard way from one trial to another. So we started UPICT. The purpose was to try to get more standard protocols like ACRIN (The American College of Radiology Imaging Network) has already developed. We needed to find some outside entity to support UPICT. So we talked to the American College of Radiology (ACR) at that time and they said they would support it because it was logical because they were the sponsors of ACRIN. Later on when I left NIH and started QIBA with the RSNA, it seemed logical to merge UPICT with QIBA. ACR agreed. So then we took UPICT and made it part of QIBA but they are similar and there is some confusion in people's mind about what is the difference between QIBA and UPICT. The way that I think about it is that in the QIBA profile you cannot deviate from the specifications; otherwise you won't be able to meet the measurement Claims that describe the levels of bias and precision that you can obtain by following the Profile. You have to do what the profile says. In UPICT, there's no measurement specified so you can modify the protocol details in the UPICT document if you need to for your particular trial.

The purpose of a UPICT document is to have everybody in a clinical trial doing the same thing for consistency. UPICT doesn't say what measurement accuracy you will achieve. And for clinical trials, as we said previously, there can be different

*11 Kurihara C. Report of the participation in the U. S. Society of Nuclear Medicine Clinical Trial Network Summit meeting. *Rinsho Hyoka (Clin Eval)*. 2010; 38(3): 623-8. Available from: http://cont.o.oo7.jp/38_3/p623-628.pdf

standards for an early phase trial versus a late phase trial. In an early trial you may actually have to be very rigid, but in a later phase trial you might allow for less rigidity because the trial is much larger and there will be more data to deal with variability. So you can change a UPICT document depending on the circumstances whereas you cannot change a QIBA Profile.

Q Thank you so much for your valuable talk. We appreciate all of your important information and wish to promote our activity of standardization in Japan more collaborating with QIBA.

(Published February 20, 2017)

*　　*　　*