Session 1: *Impact of medical metrology and standards in the clinic*

Slide 1: Title slide

Thank you, Dr. Anderson. I appreciate the opportunity to participate in this symposium.

Slide 2: Disclosures

I have no relevant disclosures.

Slide 3: Outline

The organizers asked me to review some basic concepts for the benefit of audience members who are not directly involved in some of the healthcare topics being discussed today and then review examples where medical metrology standards are needed in the clinic.

Slide 4: Why are we here?

First a statement of the problem. The short version is that clinical decision variability across healthcare is a huge problem. The setting I'm most familiar with is the wide variation inherent in the subjective, qualitative interpretation by radiologists. And although this panel 1 focuses on clinical imaging, many of the concepts I'll mention apply to other aspects of precision medicine. This comment is not limited to radiologists. These two components of performance contribute to variability in virtually all procedures and decisions made by physicians and technologists across healthcare.

Slide 5: Pancreas study

There have been hundreds of studies over the past century documenting extensive radiologist variability and I'm showing only one of them to make a couple of points. I'll skip over the study details, but note the authors' comment that "personal preference and/or opinion of the individual radiologists were responsible for 83% of the recommendation variation".

[Advance]

This is consistent with the understanding that the subjective interpretations by radiologists reflect a combination of their individual skill and value judgments. Both of those components are very resistant to standardization, which is a main reason why extracting objective, quantitative results is an enormously important path toward reducing variability. This comment is not limited to radiologists. These two components of performance contribute to variability in essentially all procedures and decisions made my physicians and technologists across healthcare

Slide 6: Editorial

And my reason for choosing this study out of hundreds is to quote from an accompanying editorial:

[Advance]

"individual patients, referring physicians, and society as a whole cannot possibly accept this degree of variability."

That study and hundreds like it document variability across groups of radiologists or groups of patients, but what are the implications for an individual patient?

Slide 7: MRI Spine

In 2017 an orthopedic surgeon in New York City who was very frustrated by the inconsistency of MRI reports from neuroradiologists, recruited one of his patients, a 63-year-old woman with low back pain and right leg radicular symptoms, to have an MRI spine scan every other day at 10 different hospitals in New York City. The surgeon had obtained money from philanthropy to pay for the scans, so insurance approval was not an issue, and the hospitals did not know they were being evaluated in this way Thus, over the course of three weeks she had 10 MRI scans.

[Advance]

The authors concluded “that no interpretive finding was reported unanimously by the radiologist at all centers and 1/3 of all reported findings only appeared once across all 10 study examination reports, which indicates that there was a significant prevalence of interpretive errors. In particular, there was a High false negative rate of 47% missed disc herniations; (which was presumably the main source of the patient’s symptoms) … and also, a High false positive rate.”

[Advance]

The authors continued: “a patient should expect to receive the same radiological diagnosis regardless of which imaging center he or she visits or which radiologist reviews the examination. Based on our extensive clinical experience the authors believe this assumption is not correct and that it can negatively impact patient care, outcomes, and costs.

***Slide 8: Thought experiment*.**

For the purpose of today's discussion consider this thought experiment:

[Advance]

Imagine that the radiologists at each of these 10 hospitals had a suite of AI tools for the spine which automatically provided the radiologist with a display of diverse quantitative information as proposed here. If the radiologists had all this quantitative information in front of them, do you think that those ten reports would vary to the extent that they did? It seems unlikely. This is technically possible today. Why, then, does this problem exist? There are several contributing impediments, and we’ll hear about some of them today.

[Advance]

One major problem is the difficulty of creating very large training data sets of standardized, population-relevant, high-quality scans. This is easier said than done, and requires resources.

Slide 9: Outline

Now, turning to some of the basic concepts.

Slide 10: Precision Medicine

The assumption underlying the concept of Precision Medicine is that each patient has a unique combination of genotypic, epigenetic and phenotypic profiles for his or her disease. In other words, two patients diagnosed as having lung cancer really have different diseases at the molecular level, or 2 patients with Type 2 diabetes are different at the molecular level.

[Advance]

Precision medicine means tailoring the diagnosis and treatment to the individual patient using patient-specific biomarkers rather than population averages.

Slide 11: Biomarkers

About 25 years ago, coincident with the emerging emphasis on precision medicine, the term biomarker was re-defined as a characteristic that's objectively measured and evaluated as an indicator of normal or abnormality. Thus, ideally, precision medicine relies on measurements. However, the term biomarker is sometimes used generically and in some contexts may not always imply a measurement.

Slide 12: What is an Image?

Before we move on to imaging biomarkers, consider these perspectives.

[Advance]

An image is an N-dimensional data set, where N is greater than one. It's at least two for anything we call an image, and for clinical imaging it's usually 3 or more.

[Advance]

Each voxel in a digital clinical image provides spatially and or temporally localized information about some property of matter with which the energy has interacted. This is critically important to understand. The voxel numbers carry chemical or biological or physiological information. This is what makes quantitative imaging an assay.

[Advance]

The process of extracting the subset of data we're interested in, that is, the subset of voxels that correspond to the normal or abnormal structures of interest, is generically referred to as segmentation. A pivotal step.

[Advance]

And because of the molecular complexity of living systems, images are often considered to be noisy, blurry datasets.

Slide 13: CT Scan

Here's a brief visual example for the benefit of those not familiar with digital clinical imaging - a lung cancer screening CT scan from several years ago.

[Advance]

There is a small nodule in that yellow box.

[Advance]

These are the Hounsfield unit values in each of the voxels in that area, and the task of an AI tool would be to identify the voxels corresponding to that nodule. It would be these that have low numbers close to 0 (which correspond to water density). But where, for example, exactly is the edge? This is one of the challenges for AI segmentation tools, that is, to correctly identify the voxels we're interested in. A second major concern is how accurately do these voxel values portray or reflect the biological or physiological information we're interested in?

Slide 14: Definition of QI

We're using the term "quantitative imaging" a lot today and so here is a formal definition.

[Advance]

I won't read all these words, but it refers to the entire process of acquiring and then extracting accurate and precise image-derived measurements.

Slide 15: QIB

Combining the definitions of biomarker and quantitative imaging, we get the definition of a quantitative imaging biomarker as a characteristic extracted from an in vivo image and objectively measured.

An essential concept to understand here is that a quantitative imaging biomarker should be thought of as an imaging assay, analogous to laboratory assays of other analytes

Slide 16: Metrology

One of the key fundamentals in metrology is this NIST statement that “a measurement result is complete only when it's accompanied by a statement of the associated uncertainty”. Many clinicians are attuned to this, because they usually see confidence intervals around blood test results, but radiologists are not accustomed to this. This is one of the educational hurdles that must be addressed.

[Advance]

Many of the generally accepted laboratory assay technical parameters apply to quantitative imaging biomarkers, but the two most important to characterize for imaging biomarkers are bias and precision.

Slide 17: Bias & Precision

To refresh your memory about these two terms, bias is how close repeated measurements are to truth (to the extent that truth is known) and precision is how close repeated measurements are to each other. One of the challenges in biological measurements is that we often don't really know truth because you can't, for example, take a tumor out of the body and get a measurement that would be the same as what it was in vivo. Thus, we often need to use reference objects, or phantoms, and we’ll hear more about that later.

Bias and accuracy ae sometimes used interchangeably. However, metrologists prefer the term bias because “accuracy” is sometimes used with a less rigorous definition.

In AI there is also a concept of bias, but it has a different definition than this metrological definition.

Radiomics are features extracted from *in vivo* images by AI tools. However, they do not have a basis in the real world. Therefore, there is no "truth", and consequently the concept of bias is usually considered meaningless in the context of radiomics. On the other hand, repeatability can be tested in radiomics, so the concept of precision does apply. Radiomics are sometimes referred to as quantitative imaging biomarkers, because they are numbers and are derived from numbers in the image. But because there is no truth for radiomic features, it is controversial as to whether they should be considered "quantitative" in the metrology sense.

Slide 18: Precision

In imaging we're interested in two subcategories of precision: repeatability and reproducibility. Repeatability is the precision that occurs when there are identical or near identical conditions between repeat measurements, and reproducibility is the precision when some of the measurement conditions change. This is common, for example, when a patient comes back for follow-up scans. In an afternoon panel session, I will show you some specific data to illustrate this.

Slide 19: Outline

Now I'll turn to some areas of clinical priority.

Slide 20: Federal Initiative

In 2015 a federal initiative on precision medicine was announced.

[Advance]

A list of disease areas of interest emerged from this initiative. To ground today's discussions around clinical needs, I used this list to search whether clinicians still seek objective biomarkers for these diseases or whether that need has been met. And because of limited time my focus was to look specifically at quantitative imaging biomarkers.

I'll show some recent publications where clinicians call for imaging measurements or improved standardization of measurements. In all the screenshots I've placed a red oval around the year; to emphasize these are current opinions.

Slide 21: Oncology

Slide 22: RECIST

In oncology the most common measurement system for over 25 years has been linear measurements of tumors called the RECIST system. Although it's been around for a long time, measurement variability is still a persistent problem. This paper from just a few months ago reviewed the literature and found continued, considerable interobserver variability.

[Advance]

They note this is primarily driven by three factors: measurement methods (whether manual versus automated), lesion selection, and imaging quality.

[Advance]

The authors state that standardizing these practices will optimize oncology patient outcomes.

Slide 23: SUV

PET scanning is also used a lot in cancer. There are many papers about reproducibility challenges in PET, but I chose this one because it describes a step in the right direction. This international benchmark is a publicly available standard database of cases against which methods can be tested and,

[Advance]

As the title says, it's a milestone towards clinical implementation. We need many more such benchmarks in the world of standardization. But keep in mind that it takes resources to create such databases for benchmarks. Those of you who have ability to allocate resources should prioritize the development of benchmark datasets. And in situations where there would be a fee to access the benchmark, there must be incentives to make it worthwhile for users to pay the fee.

Slide 24: Brain disorders

Brain Disorders is a term used collectively for psychiatric and neurologic disorders

Slide 25: Psychiatric

One of the clinical areas in dire need of objective biomarkers is the category of disorders of thoughts, emotions, and behaviors. These would correspond to the psychiatric diagnoses of psychoses, mood disorders such as depression, and obsessive-compulsive disorder. The key thing to note here is these signs and symptoms are all very subjective. It's what the patient describes. There are virtually no objective biomarkers for these disorders, so it's an enormously difficult diagnostic problem.

[Advance]

The global burden and societal impact of these disorders is huge.

Slide 26: DSM

The standard system of psychiatric diagnosis for the last several decades has been the DSM criteria,

[Advance]

but this has been criticized for years, including by NIMH, as unscientific and subjective.

[Advance]

In recent years there has been a lot of work developing more objective methods, one of which is referred to as HiTOP. These systems include many neuroimaging biomarkers and a call for better evaluation of quantitative imaging biomarkers.

Slide 27: QIB for psych

A couple of recent papers supporting this approach: one from 2024 entitled "Bringing imaging biomarkers into clinical reality in psychiatry" and

[Advance]

one from 2026 discussing the need for neuroimaging biomarkers for this category of disorders.

Slide 28: ADNI

One example from neurology, to make a different point. The Alzheimer's Disease Neuroimaging Initiative was founded in 2004, specifically with an open data model, that is, all imaging and clinical data were made available to researchers worldwide without embargo. ADNI has been going on for 20 years and has had many successes. But my reason for bringing up this mature open data model is to point out one of the challenges of reproducibility.

Slide 29: ADNI PET Reproducibility

This 2024 perspective article is from a group looking at the PET data analysis.

[Advance]

These authors wrote, "we tried to reproduce the processing pipeline of a large public data set from the ADNI study."

[Advance]

This is a summary statement, not the author's own words, indicating they hit undocumented preprocessing steps, proprietary in-house software, and outdated tools that they couldn't validate against.

[Advance]

They concluded "promoting the development of standardized pipelines and tools should be a top priority".

Slide 30: Cardiovascular disorders

Slide 31: Quantitative CV imaging

Cardiologists have extracted quantitative measures from clinical images for more than 50 years and every imaging modality.

[Advance]

However, systematic attention to assessing bias and precision in cardiovascular imaging has been a more recent focus, by which I mean the last 10-15 years.

[Advance]

In part because the thresholds between normal and abnormal are continually revised due to the expanding number of therapeutic options available in cardiology. Thus, precise quantitative imaging results are increasingly important in choosing among the therapeutic alternatives for a given cardiac condition.

Slide 32: CV Standardization efforts

I won't belabor this and just point to recent sample papers from echocardiography, CT, PET and MRI, all calling for more attention to standardization to improve precision of quantitative results.

Slide 33: infections

Slide 34: QI in Infection

This quote is from a 2024 overview article from infection experts: "understanding the host response to infection - not just detecting the pathogen - gives more accurate, timely, and actionable diagnostic information.

Slide 35: QI Tuberculosis

One important role is that imaging can be an alternative to biopsy for these difficult infections, and another is it can quantify drug distribution.

[Advance]

Tuberculosis remains the leading infectious disease worldwide.

[Advance]

There is strong quantitative evidence behind PET/CT for detecting treatment response earlier than anatomic imaging and predicting relapse risk.

[Advance]

Yet, the World Health Organization has not endorsed routine PET/CT for TB management, citing cost, access and standardization gaps.

Slide 36: Metabolic disorders

Slide 37: MASLD

In the category of metabolic disorders, I want to focus on fatty liver disease, now called MASLD and MASH.

[Advance]

Huge global problem, linked to obesity and poor dietary choices or options.

[Advance]

Significant to our discussion today is that a drug called resmetirom to treat MASLD was approved in the US in March of 2024.

[Advance]

Many payers, such as Blue Cross Blue Shield and others, now require quantitative MRI-PDFF and MR elastography for drug eligibility. MRI-PDFF is a measure of fat in the liver and MRE is a measure of stiffness which is a surrogate measure for fibrosis. These two tests together make it unnecessary to do a liver biopsy.

[Advance]

However, payer criteria for these quantitative imaging biomarkers are not uniform and there are not yet generally accepted QA or calibration criteria.

Slide 38: Liver QIBs

Two supporting papers are shown here. One from the imaging community

[Advance]

And one from a large hepatology organization promulgating guidelines for these imaging-based biomarkers.

[Advance]

I want to emphasize the importance of this MASLD-resmetirom example: this is a prototype, precedent-setting example of imaging as an assay. The quantitative imaging results are being used in the same way that blood lipid panel results would be used to determine eligibility for cholesterol-lowering drugs. All the concepts that have been developed over decades for laboratory assays for fluid and specimen biomarkers must be adapted and applied to imaging assays.

Slide 39: Other Inflammatory disorders

Slide 40: Inflammatory conditions

A list of other inflammatory diseases is long, and I won't go through all of these, but I'll mention examples from the top two on the list.

Slide 41: Ultrasound for inflammation

One validates the use of ultrasound to quantitate synovitis in rheumatoid arthritis. The measurement was synovial hypertrophy.

[Advance]

And this review of inflammatory bowel disease management recommends that ultrasound measurements of bowel wall thickness is a quantitative indicator of disease.

[Advance]

They recommend serial quantitative ultrasound measurements for response assessment, and they note that the agreement between subsequent serial measurements and baseline is 92%.

Slide 42: Summary

So, to conclude, What's broken?

[Advance]

Variation in clinical practice results in poorer outcomes and higher cost. That's the basic premise.

[Advance]

Treating physicians need quantitative results. They want measurements instead of ambiguous qualitative descriptors.

[Advance]

Those of you participating in or listening to this symposium today have the power to fix these problems.

Slide 43: Thank you.