

Clinical Working Group on Metrics Characterising Central Image Acquisition & Analysis

Helen Young, PhD AstraZeneca

James J. Conklin, MD ICON Medical Imaging

DIA Conference: MEDICAL IMAGING CONTINUUM
Path Forward for Advancing the Uses of Medical Imaging in the
Development of New Biopharmaceutical Products, Bethesda, MD
October 2, 2008

Clinical Imaging Metrics Background

- Central image collection & analysis integral to clinical trial execution
- Process & designs in place have delivered high quality data
- CRO / Pharma partnerships are a key factor to success
- Debate on-going on design, best practice and quality measures
 - Last years DIA meeting focused on common lexicon & best practice
- Lack of consensus on what is a good metric & how they should be used
- Metrics are not a replacement for:
 - Good communication and working relationships
 - High quality training, documentation and process management
 - High quality site interactions, qualification, data collection & QA
 - Good read design and reader training
- Metrics in themselves are not a measure of clinical trial success

Bruce Hillman
 Mike Morales
 Annette Chan
 DeJane Hussey
 Ria Lopez
 Robert Vaccaro
 Helen Young
 Barbara Costantini
 Colin Miller
 Dawn Flitcraft
 Jennifer Minko
 Don Cooper
 Sam Dranoff
 Kees Groenhout
 David Rauh
 Jim Conklin
 Ted Gastineau
 Ciaran Cooper
 Josh Longacre
 Gregory Lange
 Lewis Cohen
 David Mozley
 Debbie Walton
 John Griffin
 Scott Sawicki
 Alaaddin Akkaya
 Barbara Chandler
 Joanna Hicks
 Kate Stumpo
 Kevin Jaynes
 Sandra Chica
 Subashini Chandrasekaran
 Craig Lipset
 David Raunig
 K Shamsi
 Rick Patt
 Bob Ford
 Kristin Borradaile
 Kristine Szabo
 David Herron
 Joyce Suhy
 Mark Tengowski
 Christina Mastandrea
 Patrick Chokron

ACR Image Metrix
 ACR Image Metrix
 Amgen
 Amgen
 Amgen
 Amgen
 AstraZeneca
 Bio Imaging Technologies
 Bio Imaging Technologies
 Bio Imaging Technologies
 Bio Imaging Technologies
 Biomedical Systems
 Biomedical Systems
 Cardialysis
 Eli Lilly
 Icon Medical Imaging
 Icon Medical Imaging
 M2S
 M2S
 M2S
 Medarex
 Merck
 Merck
 Novartis
 Novartis
 Perceptive Informatics
 Perceptive Informatics
 Perceptive Informatics
 Perceptive Informatics
 Perceptive Informatics
 Perceptive Informatics
 Perceptive Informatics
 Pfizer
 Pfizer
 Rad-MD
 Rad-MD
 RadPharm
 RadPharm
 Schering-Plough
 Synarc
 Synarc
 Virtual Scopics
 WorldCare Clinical
 WorldCare Clinical

- Imaging CROs & Pharma
 - Experience covering all aspects of clinical process
- Representation across skills
 - Biostatistics
 - Medical Writing
 - Operations
 - Imaging Science

- Working group objectives for 2008
 - Business case for clinical metrics & how they should be used
 - Review of the clinical process & application of metrics
 - Identify issues for wider debate & consultation
 - Consult with other MMC working groups
 - Develop & interrogate draft metrics in the MCC format
 - Enable peer review of the draft metrics
 - Finalise and present version 1.0 metrics in 2009.
- Regular monthly meetings & ad hoc activities

How should metrics be used ?

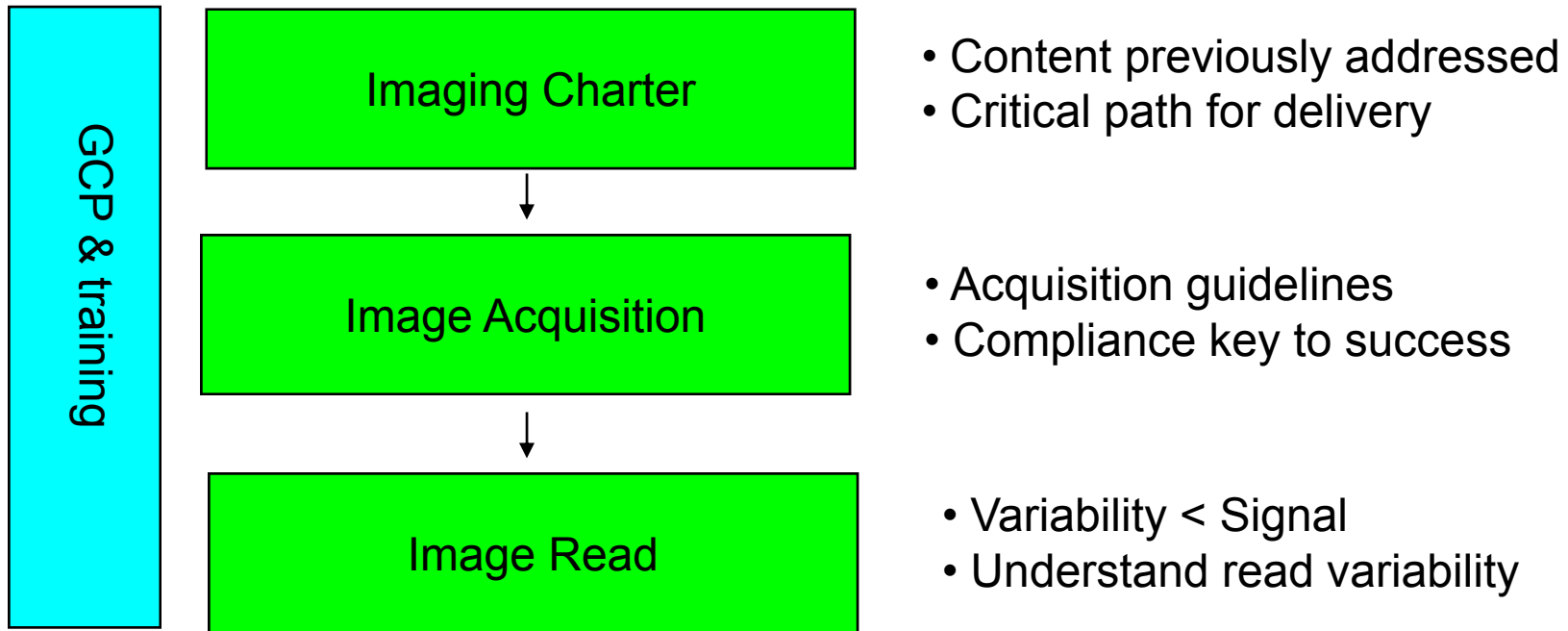
- Tools for Pharma/CRO partners
 - Quantify “time to” and “performance” for process elements
 - Assist with project planning & management
 - Identify issues for more in depth analysis
 - Bench mark process elements

What is the impact of measurement ?

- Help process improvement & working relationships
- Increase scientific/clinical understanding
- Drive process improvement (e.g. training needs)

What are the potential pitfalls ?

- Used without understanding / exploration of drivers
- Focus on metric rather than education/training
- Poor understanding – misuse as a measure of clinical trial success



- Compliance & training critical to success of central review
- Metrics would be useful for Imaging charter, acquisition & read
- Assessment of read variability is complex & requires statistical input
- Consultation with MCC ECG clinical group

- Imaging charter development time
- Acquisition guideline amendments
- Reader variability categorical assessments
- Reader variability continuous measurements

Adjudication referred for wider debate at this meeting

Metric	Category	Metric Title	Definition	Formula / Example	Unit of Measure	Reporting Frequency	Target	Business Driver(s) / Benefit Statement
Clinical #1	Independent Review Charter	Time to develop and write independent review charter	<p>Minimum: time from when Imaging CRO receives study protocol <u>and</u> start date agreed with Sponsor to the date v1.0 is signed</p> <p>Additional analysis on a "for cause" basis:</p>	<p>Formula: (date v1.0 is signed minus date agreed with CRO/sponsor for charter start)/ 7</p> <p>Specific Example: (30JUN08 - 25APR08) = 66 days/7 = 9.4 weeks</p>	weeks		N/A	Knowing the time needed to develop and write an Independent Review Charter is essential for project planning on the part of the Sponsor and the Imaging CRO. The Charter must be finalized before other systems can be put in place and before reviews may begin. Metric can be utilized to assess relationship, communication and responsiveness for CRO/PhRMA partners.
Clinical #2	Acquisition Protocol Robustness	Technical Acquisition Guideline Amendments	<p>Minimum: number of image acquisition technique-related amendments to the site acquisition guidelines per modality for a study</p> <p>Additional analysis on a "for cause" basis:</p>	<p>Formula number of issued deviations or amendments to the site acquisition guidelines per modality for a Study. Tracked across studies for process improvement</p> <p>Specific Example:</p>	number	per study	0	This metric is particularly relevant when introducing or monitoring new image acquisition guideline performance across multiple sites and vendor platforms. It allows both the sponsor and CRO to understand robustness of acquisition protocol and drive change in response to technical non compliance generating more generic and useful acquisition protocols.

- Business Case
 - Knowing the time needed to develop & write an Independent Review Charter is essential for project planning.
 - on the part of the Sponsor and the Imaging CRO.
 - The Charter is often submitted as part of regulatory process & ideally should be in place prior to first patient entered
 - The Charter must be finalized before other systems can be put in place and before reviews may begin.
 - Metric can be utilized to assess relationship, communication and responsiveness for CRO/PhRMA partners including sponsor responsiveness

Metric Title

Time to develop and write independent review charter

Definition

Time from when Imaging CRO receives study protocol and Start date agreed with Sponsor to the date v1.0 is signed

Formula

(date v1.0 is signed minus start date agreed with sponsor) / 7

Two sources of image quality issues

- Method inadequate e.g.
 - Inadequate specification
 - Platform specific parameters do not generalize
 - Normal volunteer methods do not translate
 - Solutions in amendments to acquisition guidelines
- Site non-compliant with protocol
 - Solutions in training and awareness

MCC Operational Work Group is developing metrics to assess occurrence rate for Image quality issues

Acquisition Guideline Amendments

- Particularly relevant when introducing or monitoring new image acquisition guideline performance across multiple sites and vendor platforms.
- It allows both the sponsor and CRO to understand robustness of acquisition protocol and drive change in response to technical non-compliance generating more generic and useful acquisition protocols

Metric Title

Technical Acquisition Guideline amendments

Definition

Number of image acquisition technique-related amendments per modality per protocol

Track data over time to assess process improvement

Reader Variability Metrics – business case

- Variability measures assists both the CRO and sponsor in understanding the reproducibility of an imaging end point
 - with respect to intra and inter-reader variability
 - with potential impact on study design and read design
- For a robust end point, the expected change in signal is much greater than variability due to methodology or biological variation

Reader Variability Metrics – business case

- Caveats are that the causes of variability need to be investigated thoroughly before drawing conclusions including those around read quality or individual reader performance.
- These metrics could in certain circumstances be used to identify readers who had lower reproducibility compared to other readers on the study and steps taken to understand and remediate.

Observer Variability Definitions

Draft lexicon reviewed at 2007 meeting

Inter-Observer Variability or Inter-Reviewer Variability	The variability in the interpretation of a set of images by different reviewers.
Intra-Observer Variability or Intra-Reviewer Variability	The variability in the interpretation of a set of images by the same reviewer after an adequate period of time inserted to reduce recall bias.

Acknowledgment to David Raunig for statistical contribution to development of proposed reader variability metrics

Categorical Variable Definition

- Measurement scale defined by a distinct & relatively small set of categories
- Categorized by a quantitative threshold or a qualitative diagnosis/evaluation
- Ordinal – Natural ordering usually between more than 2 states
 - Progressive Disease → Stable Disease → Responsive Disease
 - Absent → Mild → Moderate → Severe
- Nominal – No natural ordering
 - Biogenic / Environmental / Combination
 - Oriented / Disoriented
 - Survive / Death

Categorical Variable Definition

- Data are measured along a scale that includes a large number of values
- The relationship between values has a clear mathematical meaning
- Examples are:
 - RECIST values (not objective response criteria e.g. sum LD)
 - SUV
 - Number of lesions
 - Scores (eg. 1,2,3,...) can be treated as continuous

- Appropriate for inter & intra reader variability assessment for categorical variables
- A single index metric that compares the probability of agreement to that probability expected by pure chance.

$$\kappa = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

– Weighting

- Categories that are closer the natural order have more weight for agreement than agreement between categories that are more separated
- Examples for 2 readers
 - High weighting → PD-PD, SD-SD and PR-PR pairs ($W = 1$)
 - Medium weighting → PR-SD and PD-SD pairs ($W=.75$)
 - Low weighting → PD-PR pairs ($W=0$)

Kappa Example

2 Readers / 4 Categories

		Reader 1 Categories									
						1	2	3	4		
Reader 2 Categories		1	2	3	4	Column Sum	Quadratic Weights				
x	1	52	40	18	2	112	0.89	0.56	0.00	-0.78	
	2	15	35	10	5	65	1.00	0.89	0.56	0.00	
	3	15	16	24	15	70	0.89	1.00	0.89	0.56	
	4	3	7	21	43	74	0.56	0.89	1.00	0.89	
Row Sum		85	98	73	65	321					
x/N	1	0.162	0.125	0.056	0.006	0.349	N=321				
	2	0.047	0.109	0.031	0.016	0.202					
	3	0.047	0.050	0.075	0.047	0.218					
	4	0.009	0.022	0.065	0.134	0.231					
Row Sum		0.265	0.305	0.227	0.202	1.000					
x*f	1	0.144	0.069	0.000	-0.005	0.208	po=sum(x*f)				
	2	0.047	0.097	0.017	0.000	0.161	0.762				
	3	0.042	0.050	0.066	0.026	0.184					
	4	0.005	0.019	0.065	0.119	0.209					
Row Sum		0.237	0.235	0.149	0.140	0.762					
Ex	1	0.092	0.107	0.079	0.071	0.349					
	2	0.054	0.062	0.046	0.041	0.202					
	3	0.058	0.067	0.050	0.044	0.218					
	4	0.061	0.070	0.052	0.047	0.231					
Row Sum		0.204	0.235	0.175		0.614					
Ex*f	1	0.082	0.059	0.000	-0.055	0.086	pe=sum(Ex*f)				
	2	0.054	0.055	0.026	0.000	0.134	0.597				
	3	0.051	0.067	0.044	0.025	0.187					
	4	0.034	0.063	0.052	0.041	0.190					
Row Sum		0.221	0.243	0.122	0.011	0.597					
k=(po-pe)/(1-pe)		0.409									

- Rules of Thumb
 - Slight Agreement $k < 0.2$
 - Fair Agreement $0.2 < k \leq 0.4$
 - Moderate Agreement $0.4 < k \leq 0.6$
 - Strong Agreement $0.6 < k \leq 0.8$
 - Outstanding Agreement $k > 0.8$

Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*.1977;33(2):363-74.

- A measure of the linear relationship between 2 continuous variables

$$\rho = \frac{\text{covariance}(x, y)}{\sqrt{\text{var}(x) * \text{var}(y)}}$$

- Assumes a linear relationship regardless of true relationship
- Not a measure of agreement
 - Does not account for y-intercept
 - Does not account for slopes different from 1

Concordance Correlation Coefficient

- Definition
 - Correlation between 2 variables corrected for
 - Mean differences
 - Slope $\kappa 1$
 - Variability in the X-variable (Interchangeability of X and Y)

$$ccc = \frac{\text{var}(x) + \text{var}(y) - \text{var}(x - y)}{\text{var}(x) + \text{var}(y) + (\text{mean}(x) - \text{mean}(y))^2}$$

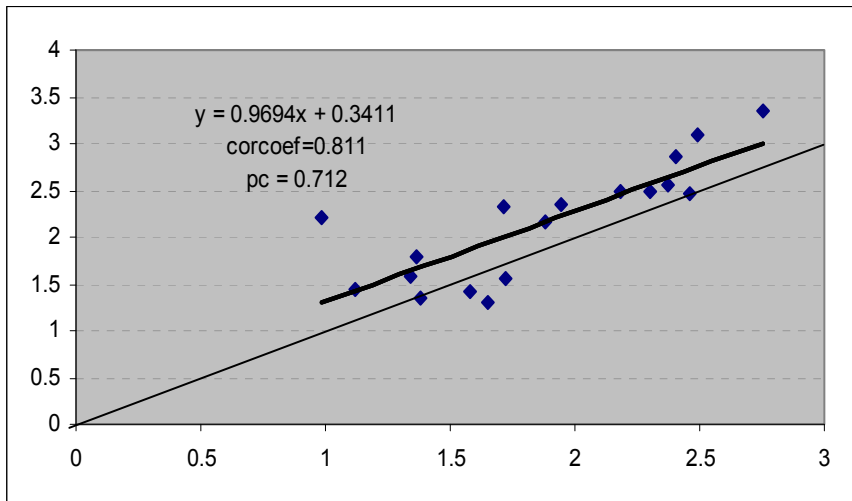
- Proposal- use the concordance correlation coefficient as a measure of agreement between two methods (or readers for continuous measures).
 - Differentiates between linearly related and agreement
 - Accounts for offsets or biased results

Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45:255-268.

Correlation Coefficient Examples

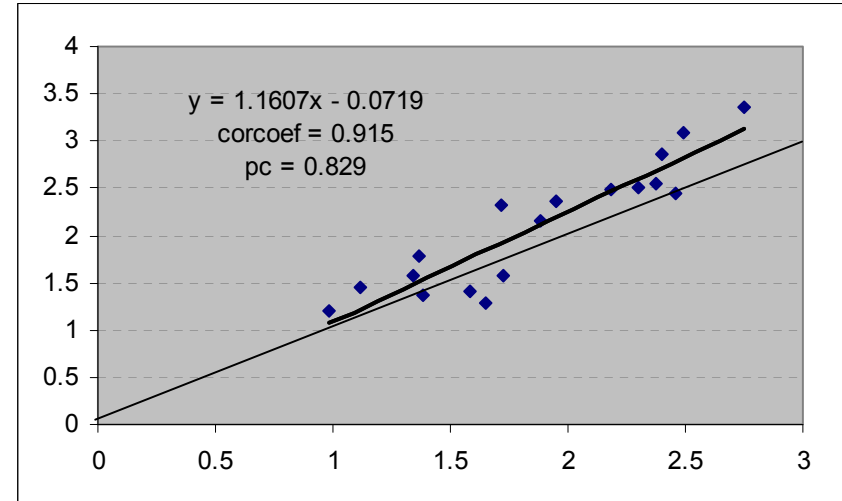
Correlation Coefficient

Intercept > 0
Slope = 1



Continuous Correlation Coefficient

Intercept = 0
Slope > 1



Concordance Correlation Coefficients correct for:

- deviation from identity slopes
- non-zero y-intercepts

Statistical Significance & Sample Sizes

- Please engage Biostatistician when using statistical metrics
 - Experimental design, application & significance of findings
- Kappa and sample size
 - Function of square of the average Kappa for two experiments
- A general guideline would be $n = 60$ for two categories
 - 0.4 to 0.8 with 80% confidence, one sided significance of 0.2
- Sample sizes go down with multiple categories
- For our worked example $n = 40$
 - 0.4 to 0.8 with 80% confidence, one sided significance of 0.2
- Sample sizes go down again for continuous variables $n = 15$
 - The samples size for same decrease in correlation coefficient is 15
 - For small samples may be better to use underlying measurement e.g. Sum LD, SUV etc rather than response category.

- Kappa popular & well-developed single metric for agreement for categorical
 - Can use weights though with caution
 - Can be very sensitive to the numbers in each class
- ICC & CCC are popular & well-developed single metrics for agreement for continuous measurements
 - CCC can account for scale and shift differences in readers
- Single indices of agreement do not tell the whole story
 - Underlying causes of discrepancy should be examined
- Statistical Metrics applied with caution & understanding
 - Operationally feasible sample sizes for Kappa may be small (not statistically sig)
- Possibility for benchmarking & understanding read variability
- May lead to process change increasing precision & accuracy of read method

- Virtually every client requests variability measures!
 - Inter-reader variability
 - Intra-reader variability
 - Adjudication rate (AR)
- No standard definition of metric or threshold!
- Adjudication Rate variable and highly dependent on denominator
 - Subjects, images, lesions, timepoints

- Kappa calculated for nearly every diagnostic imaging trial using **categorical** data
 - K can be used for 2 or more outcomes

Can Kappa be used for a measure of reader variability?

- Correlation coefficients routinely calculated in trials using **continuous** data
 - Differentiates between linearly related and agreement
 - Accounts for offsets or biased results

Can concordance correlation coefficient be used to measure reader variability?