MULTIPARAMETRIC QUANTITATIVE IMAGING BIOMARKERS

A FRAMEWORK FOR ESTIMATING AND TESTING TECHNICAL PERFORMANCE

Quantitative Imaging Biomarkers Alliance (QIBA) Multiparametric Metrology Group

Quantitative Imaging Biomarkers Alliance RSNA

QIBA METROLOGY PAPERS

Quantitative imaging biomarker (QIB): "objectively measured characteristic derived from an in vivo image as an indicator of normal biological processes, pathogenic processes, or response to a therapeutic intervention" (Sullivan et al 2015).

Types of validation of QIBs and tests based on them (adapted from BeST glossary):

Technical performance validation (analytic validation): assessment of technical performance characteristics of a QIB measurement protocol (e.g. test-retest repeatability, bias and linearity, reproducibility as a function of differing image acquisition and processing protocol aspects)

↓

Clinical validation: assessment of the ability of the QIB to detect or predict disease status, outcome, or phenotype of interest (e.g. diagnostic accuracy, association of QIBs with a clinically significant characteristic or molecular biomarker, association of QIBs and OS or PFS)

↓

Assessment of clinical utility: evaluation of the net improvement in patient outcome resulting from the use of the test to guide treatment decision

To provide guidance on technical performance validation of measurement processes for QIBs, the QIBA Metrology Working Groups published a series of papers in 2015:

- Terminology: Kessler et al (2015), *Statistical Methods in Medical Research* 24 (1), pg. 9-26.
- Performance: Raunig et al (2015), *Statistical Methods in Medical Research* 24 (1), pg. 27-67.
- Algorithm comparison: Obuchowski et al (2015a), *Statistical Methods in Medical Research* 24 (1), pg. 68-106.
- Case example: Obuchowski et al (2015b), *Statistical Methods in Medical Research* 24 (1), pg. 107-140.
- Meta-analysis: Huang et al (2015), *Statistical Methods in Medical Research* 24 (1), pg. 141-174.
- Overview for radiologists: Sullivan et al (2015), *Radiology* 277 (3), pg. 813-825.

→ **Concepts and methods from these papers were intended for a single QIB.**

BeST (Biomarkers, EndpointS, and other Tools Resource). ncbi.nlm.nih.gov/books/NBK338448.

MULTIPARAMETRIC QUANTITATIVE IMAGING BIOMARKERS

Many clinical scenarios involve multiple QIBs, e.g. multiparametric MRI in prostate cancer involving T2-weighted, diffusion weighted, and dynamic contrast-enhanced MRI. These multiparametric QIBs present new issues. For example,

- Approaches from the above manuscript cannot simply be applied to each QIB individually as it is likely to result in overinflated Type I error (see simulation studies example to the right).
- If data from multiple studies need to be combined to accrue a sufficiently large sample size, not all QIBs of interest may have been measured in each study.
- The QIBs often need to be combined into a score for differentiating cases of various phenotypes and predicting clinical outcomes; selecting which QIBs to use, combining them, and checking that the resulting score is meaningful in terms of phenotype or outcome requires careful application of appropriate statistical techniques.

→ **The purpose of this effort is to develop a similar set of guidelines for the multiparametric scenario to address these issues.**

Correlation between QIBs	Probability of declaring underlying change in at least one QIB
Substantial correlation (Pearson correlation about 0.6)	0.193
Minimal correlation (Pearson correlation between -0.01 and 0.01)	0.267
Mixed levels of correlation (correlations are substantial half the time and minimal half the time)	0.247

Simulation studies results. Test and retest measurements for each of six QIBs were generated with no underlying change in any of them. The proportion of times in which a change in at least one QIB was detected was estimated by comparing the magnitude of the difference between the test and retest measurements of each QIB to the corresponding repeatability coefficient (RC; i.e. threshold below which differences in repeat measurements obtained under the same conditions should fall with 95% probability). These proportions should be close to 0.05.

QIBA Multiparametric Metrology Working Group members: H. Barnhart (Duke), M. Boss (ACR), A. J. Buckler (Elucid Bioimaging Inc.), P. E. Cole (Bayer), M. L. Giger (Chicago), A. Guimaraes (Oregon Health and Science), T. J. Hall (Wisconsin), E. P. Huang (NCI), J. Kalpathy-Cramer (MGH), P. E. Kinahan (Washington), M. Kondratovich (FDA), C. Moskowitz (MSKCC), N. Obuchowski (CCF), K. O'Donnell (Canon Medical Research USA Inc.), G. Pennello (FDA), N. Petrick (FDA), L. H. Schwartz (Columbia), D. C. Sullivan (Duke). **Acknowledgements:** J. Koudelik (RSNA), J. Lisiecki (RSNA).

DEVELOPMENT OF THE FRAMEWORK FOR THE MULTIPARAMETRIC CASE

Use Case 1: QIBs are treated as a multivariate vector (not to be combined mathematically into a score or prediction).

- Inferences for bias for a QIB can be improved by incorporating information from other QIBs if the QIBs are correlated (see table to the right).
- Methodology developed to test for substantial changes in these QIBs based on the Mahalanobis distance *Q* (i.e. distance between the zero vector and the vector of changes, corrected by variance); results in correct Type I error (see figure and table to the right).
- Future work: develop statistical methodology to combine data from multiple sources when not all QIBs are available in all sources.
- Future work: develop guidelines to establish the optimal set of imaging parameters to obtain measurements of the QIBs.

Use Case 2: QIBs are combined via statistical model or decision rule to produce a classification of cases according to phenotypes (e.g. molecular subtype).

- Prediction accuracy of the model should be shown to be better than some null value (e.g. random guessing).
- Reproducibility of the score (i.e. model output) and classification should be reported.
- Future work: to develop a glossary of terms relevant to this use case.
- Future work: to achieve consensus on which aspects of the above to report and appropriate metrics associated with these aspects.
- Future work: to put forth guidelines on how to properly develop the model and assess its prediction accuracy.

Use Case 3: QIBs are combined in a similar fashion to generate a risk score or prediction of response to therapy or clinical outcome.

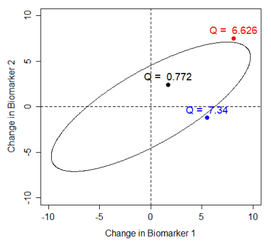
- Many concepts from Use Case 2 should also apply, but the focus on risk score, response prediction, or clinical outcome necessitates different metrics.
- Future work items similar to those of Use Case 2.

Use Case 4: radiomic analyses.

- Radiomic features may not be biomarkers in the mathematical sense, but many concepts from Use Case 2 and 3 should also apply here.
- Development of guidelines for this use case to be started at a later date.

Technical performance of first QIB as a function of the second	Truth: mean bias and wSD for first QIB	Correlation between the two QIBs	Estimates of bias and wSD of first QIB	
			Ignoring second QIB	Not ignoring second QIB
No association between the two QIBs	Bias: 0 wSD: 1.60	0	Bias: -0.025 (0.043) wSD: 1.578 (0.012)	Bias: -0.025 (0.043) wSD: 1.578 (0.012)
		0.5	Bias: -0.025 (0.043) wSD: 1.578 (0.012)	Bias: -0.025 (0.040) wSD: 1.578 (0.012)
		0.9	Bias: -0.025 (0.043) wSD: 1.578 (0.012)	Bias: -0.025 (0.033) wSD: 1.578 (0.012)
Bias in first QIB increases slightly as value of second QIB increases	Bias: 2.19 wSD: 1.63	0	Bias: 2.23 (0.043) wSD: 1.609 (0.012)	Bias: 2.23 (0.043) wSD: 1.609 (0.012)
		0.5	Bias: 2.23 (0.044) wSD: 1.609 (0.012)	Bias: 2.23 (0.040) wSD: 1.609 (0.012)
		0.9	Bias: 2.23 (0.046) wSD: 1.609 (0.012)	Bias: 2.23 (0.033) wSD: 1.609 (0.012)
wSD of measurements of first QIB increases slightly as value of second QIB increases	Bias: 0 wSD: 3.24	0	Bias: -0.026 (0.043) wSD: 3.233 (0.019)	Bias: -0.026 (0.043) wSD: 3.233 (0.018)
		0.5	Bias: -0.026 (0.043) wSD: 3.233 (0.019)	Bias: -0.026 (0.040) wSD: 3.233 (0.018)
		0.9	Bias: -0.026 (0.043) wSD: 3.231 (0.019)	Bias: -0.026 (0.033) wSD: 3.231 (0.018)

Simulation studies results. Values of two QIBs were generated using different relationships between the first and the second. Inferences on the bias and within-case standard deviation (wSD) of the first QIB were performed ignoring the second QIB and taking into account the second QIB. Standard errors of bias are improved by incorporating data from the second QIB when the QIBs are correlated.



For two QIBs, vector of changes should lie within the ellipse with 95% probability and $Q < 5.99$ if no underlying change occurred in either QIB. **Red:** underlying change in both QIBs. **Blue:** underlying change in one QIB. **Black:** no underlying change in either.

Correlation between QIBs	Probability of declaring underlying change in at least one QIB
Substantial correlation (Pearson correlation about 0.6)	0.053
Minimal correlation (Pearson correlation between -0.01 and 0.01)	0.053
Mixed levels of correlation (correlations are substantial half the time and minimal half the time)	0.053

Proportion of times in which a change in at least one QIB was detected based on *Q* for the simulation studies described previously. These proportions should be close to 0.05.