QIBA Algorithm Comparison Activities: Digital Reference Objects and Software Evaluation

Daniel P. Barboriak[1], Nancy Obuchowski[2], Alexander R. Guimaraes[3], Cathy Elsinger[4], Gudrun Zahlmann[5], Daniel Sullivan[1], Edward F. Jackson[6]
James Voyvodic[1], Edgar DeYoe[7], Jay Pillai[8], Paul Kinahan[9], Hendrik Laue[10], Thomas L. Chenevert[11], Dariya Malyarenko[11], Michael A. Boss[12]

[1]Duke University, [2]The Cleveland Clinic Foundation, [3]Oregon Health & Science University, [4]NordicNeuroLab Inc., [5]F. Hoffman - La Roche, Ltd.,
[6]University of Wisconsin, [7]Medical College of Wisconsin, [8]Johns Hopkins University, [9]University of Washington, [10]Fraunhofer-MEVIS, [11]University of Michigan, [12]National Institute of Standards and Technology

What is a Digital Reference Object (DRO)?

"Digital reference images are synthetic images that have been created by computer simulations of a target in its environment; the image acquisition device (i.e. scanner) is not involved but similar noise artifacts are added to the image. An advantage of these approaches is that the true value is known. A disadvantage of the synthetic image approach is that currently these methods are approximations to the real images and do not faithfully represent all the important subtleties encountered in real images, especially the second or higher order moments of the data (e.g. the correlation structure in the image background). Phantoms and digital reference images may be used to establish a minimum performance requirement for QIB algorithms." (Ref: Obuchowski et al., SMMR 2014)

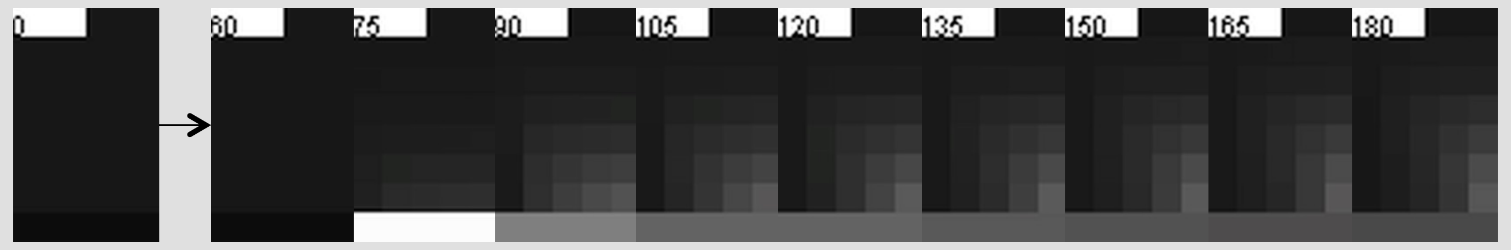
Measures of agreement: Disaggregate vs. aggregate

"There are two general approaches to evaluate the degree of closeness between measurements by an algorithm and the true value: disaggregated and aggregated approaches. In the disaggregated approach, the performance of the algorithm is characterized by two components: bias and precision. We would assert that the algorithm performs well if the algorithm has both small bias and high precision. In the aggregated approach, the performance of the algorithm is evaluated by a type of agreement index which aggregates information on bias and precision. With this approach we would assert that the algorithm is performing well if there is ''sufficient'' degree of closeness judged by the agreement index between the algorithm and the true value. If substantial disagreement is found, then the sources of disagreement, i.e. bias or precision or both, can be investigated." (Ref: Obuchowski et al., SMMR 2014)

Previous Work: NIBIB-funded Projects, Round 1-4

DCE-DRO: PI Daniel Barboriak

In DCE-MRI, DROs are created in order to evaluate software analysis packages with respect to their ability to determine T_1 using variable flip angle MRI, and to generate parameter maps of K^{trans} and v_e. The initial DCE DRO was created as part of the NIBIB-funded QIBA Round 1 projects, and recently extended through a Round 4 project. There is now a set of DCE DROs that explore a variety of conditions: in general, the DROs consist of a grid of patches, each 10x10 pixels, with combinations of K^{trans} and v_e. These patches' signal intensities are temporally evolved (over 10 minutes) using a set of parameter values (field strength, flip angle, TR, time interval and temporal jitter, intrinsic tissue and vessel T_1 and equilibrium magnetization), using a modified Tofts Kermode 2-parameter model. DICOM image sets are generated, and then used as input for a given software analysis package, allowing evaluation of a given package's performance with a known ground-truth dataset.



In the figure above, a portion of a DCE DRO (v6) is depicted, showing the first three minutes of a simulated DCE experiment. Injection of gadolinium contrast occurs at 60 seconds. v_e increases in each 10x10 patch of pixels along x {0.01, 0.05, 0.1, 0.2, 0.5}, while K^{trans} increases along y {0.01, 0.02, 0.05, 0.1, 0.2, 0.35}. The bottom row of pixels (50x10) represent the vascular region of interest, and the peak signal is recorded in all images around the time stamp in the upper-left corner, and a zero patch is located in the upper-right corner.

With the many parameters that can affect DCE, several DROs have been created to test the robustness of analysis packages under a variety of anticipated conditions, such as different field strengths, different tissue and vascular T1 values, varying equilibrium magnetization, noise, altered vascular input functions, and alternate sampling rates and temporal jitter. Further details, and the actual DCE DROs themselves are available at:

http://qidw.rsna.org/community/6

A number of existing DCE software packages were evaluated using these DROs. Packages were compared using disaggregated metrics such as mean bias and within-subject standard deviation (wSD) and aggregated metrics including root mean square deviation (RMSD), and the concordance correlation coefficient (CCC).

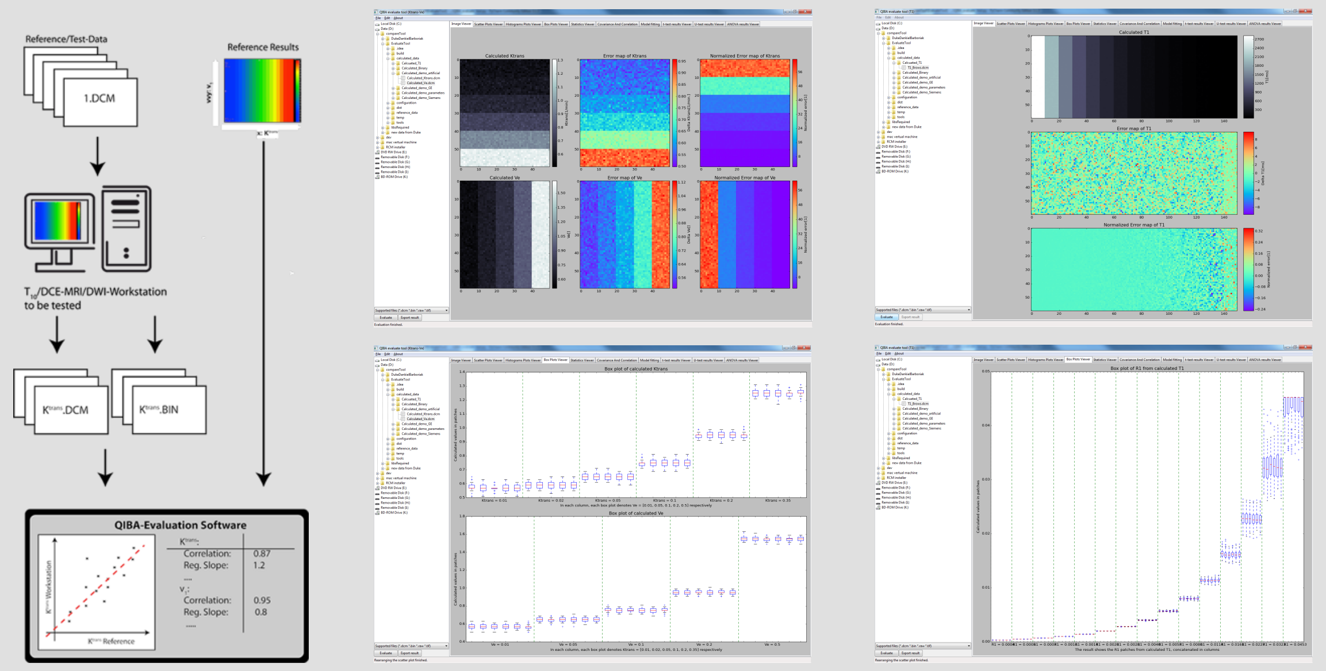
When varying tissue T_1, software packages exhibited different patterns of bias, indicated a need to evaluate software across a range of T_1 values. Bias was affected by field strength for two packages, warranting further study. The use of a reduced cardiac output vascular input function had little impact on K^{trans} bias for noise-free DROs, nor on ranking by aggregate metrics using noise-added DROs. Finally, summary recommendations indicate that aggregated metrics may be a better tool to facilitate software evaluation and comparison.

Dr. Obuchowski analyzed the utility of a specific aggregated metric, total deviation index at 95% coverage (TDI) in evaluating T_1 mapping software. It was found when a software package is ranked similarly based on bias and wSD, the ranking by TDI often agreed; when a software package is ranked differently based on bias and wSD, the ranking by TDI often agreed more closely with wSD. Based on this study, the properties and performance of TDI seemed to be an excellent choice as an aggregate measure.

DCE-DRO Software Evaluation: PI Hendrik Laue

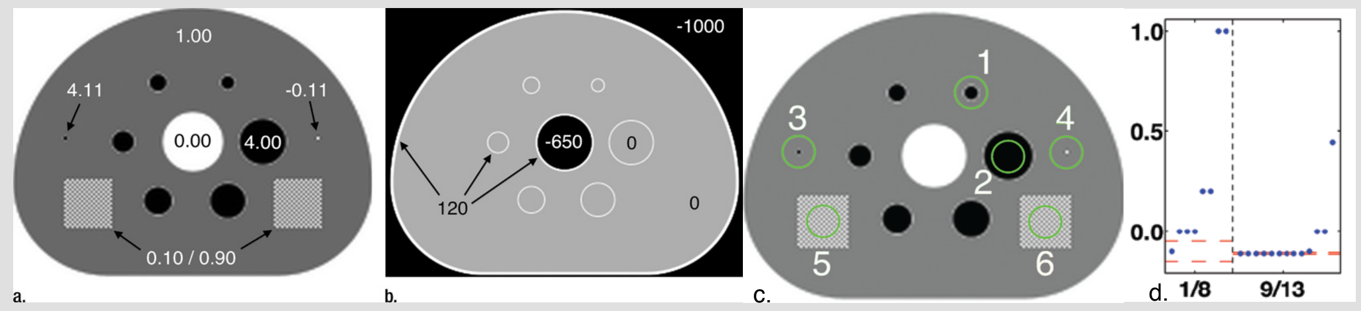
The QIBA DRO Evaluation Tool (QDET) was developed to assist in software evaluation. An open-source, Python-based software package, QDET is to designed to take in calculated parameter maps of T_1, K^{trans}, and v_e, and compare them to reference maps. QDET automatically generates error and normalized error plots, scatter plots, histograms, and box plots, and is designed to work in conjunction with the DCE DROs developed by Dr. Barboriak. Statistics from these plots are also generated in HTML format, facilitating further analysis and results dissemination.

Further details and source code can be found at: **http://qidw.rsna.org/community/6**



PET-CT SUV DRO: PI Paul Kinahan

A PET-CT DRO was created, based on the NEMA image quality physical phantom, with SUV_{BW} (Fig. a) and Hounsfield units (Fig. b) representing water, air, lung, and polymethylmethacrylate. The DRO represents an abdominal cross-section, with six spheres ranging in diameter from 10 to 37 mm, and a central 50 mm cylinder, with 1mm thick plastic walls. Isolated pixels with specific SUV_{BW} values, and checkerboard regions with alternating values were also included. 21 different PET/CT software packages at 16 sites were evaluated with the DRO. 25 mm diameter circular ROIs were drawn within the PET DRO (Figure c; a spherical ROI was used for Region 6 when allowed by software), and each site reported the maximum, minimum, mean, and standard deviation of SUV_{BW} within the ROIs.



As an example of the data collected, Figure d demonstrates the minimum SUV value recorded in ROI 4: while a large number of the analysis packages correctly determine this to be -0.11, a majority do not, with some packages reporting SUV_{min} to be 1, i.e., the background level. This particular example and others (see Publications for a manuscript with full details), demonstrate that SUV metrics are not always properly computed, and that large variability exists across software packages. While to date only SUV_{BW} has been investigated, other metrics are anticipated to exhibit similar or larger variability. Image noise was not part of this investigation; however, the controlled noise of ROIs 5 and 6 exhibited significant variability in average and max values. Based on these results, variability across sites and software is large enough to warrant concern about quantitative reproducibility, especially in multicenter trials; the DRO can serve as a standard with which to validate individual site workstations.

PET-CT DROs can be found at: http://qidw.rsna.org/community/7

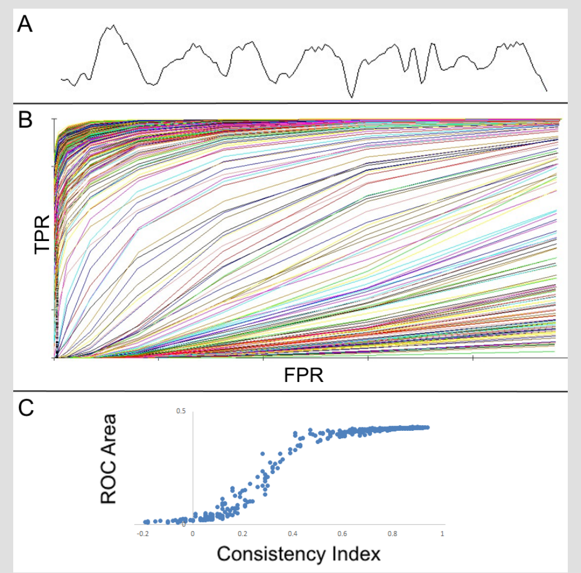
fMRI DRO: PIs Jim Voyvodic, Edgar DeYoe, Jay Pillai

For functional MRI, DRO's are generated by combining multiple independent time-varying signal components to create a time-series of realistic EPI brain images with known brain function properties. Individual signal components include:

- Static T_2^*-weighted EPI images
- Task-dependent T_2^* signal fluctuations
- Functional anatomy brain activation weighting map ("truth" map)
- Resting-state T_2^* signal fluctuations
- Task-performance weighting over time
- Head motion transformations over time
- Neurovascular uncoupling weighting map

For example, the influence of variability of task-performance was tested by generating 400 DROs that differed only in the consistency of task performance, based on empirical measures from 400 real patients (e.g., Fig A). FMRI analysis of each DRO produced activation maps that were compared to "truth" to obtain ROC curves (Fig B). Correlating the "consistency index" (CI) generated during fMRI QA preprocessing, to the areas under the ROC curve for each DRO showed that CI values can be used as a conformance constraint in our fMRI biomarker profiles (Fig C). To achieve the Claims, the CI value should be at least 0.4.

fMRI DRO datasets can be found at: http://qidw.rsna.org/community/13



Future Work: Round 5 projects

Aggregated Measures of Agreement for QIB Validation: an Open Source Toolkit , PI Daniel Barboriak

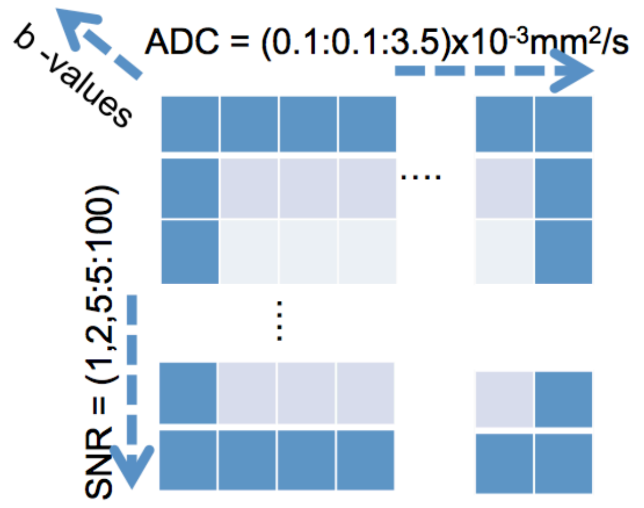
The purpose of this project is develop open source software to calculate aggregated measures of agreement in order to facilitate image analysis algorithm development, comparative analysis of algorithm output, and demonstration of technical compliance. It will build off of previous efforts to develop DCE-MRI DROs, the QDET for comparing software packages, and evaluation of various figures of merit. Focusing on aggregate measures, the project will provide open-source analysis of software performance in the form of CCC, RMS, the total deviation index (TDI), and Bland-Altman Limits of Agreement (LOA), allowing ranking of various analysis packages. By allowing inputs of nominal ground truth, means and squared errors to be used as input, the analysis could be applied to results derived from physical phantoms such as the DWI phantom, T_1 response phantom, and lesion simulation phantoms, extending the utility of the toolkit across all QIBA efforts.

DWI-DRO development for ADC analysis, PI Dariya Malyarenko:

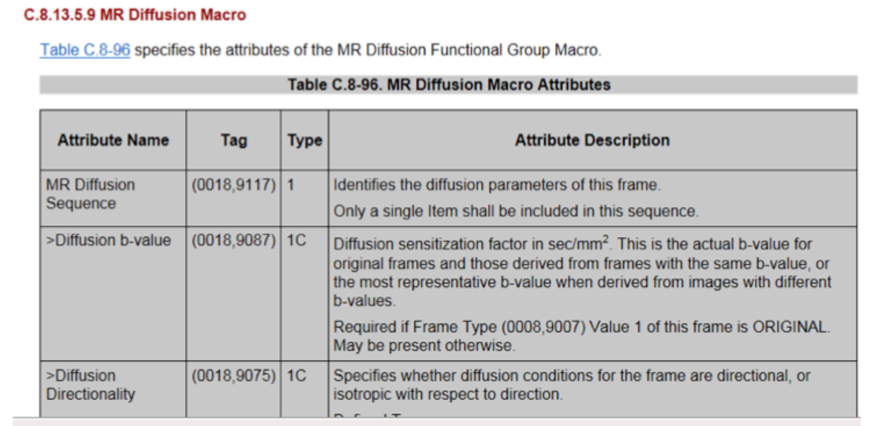
Provide DWI DRO for relevant range of tissue diffusion values and Rician noise utilizing standard diffusion DICOM attributes for SW testing and acquisition optimization.

DWI-DRO Project Deliverables:

- Definition of parameter space suitable for DRO derivation
- Definition of DICOM-compliant trace-DWI DRO attributes
- DRO DICOM generation using defined diffusion models and input parameters, including Rician noise
- Test procedures to evaluate DRO by reproducing input parameters and models
- 3D DWI-DRO standard DICOM set with analysis and performance evaluation instructions



DWI-DRO DICOM Attributes:



C.8.13.5.9 MR Diffusion Macro

Table C.8-96 specifies the attributes of the MR Diffusion Functional Group Macro.

Table C.8-96. MR Diffusion Macro Attributes

Attribute Name	Tag	Type	Attribute Description
MR Diffusion Sequence	(0018,9117)	1	Identifies the diffusion parameters of this frame. Only a single Item shall be included in this sequence.
>Diffusion b-value	(0018,9087)	1C	Diffusion sensitization factor in sec/mm². This is the actual b-value for original frames and those derived from frames with the same b-value, or the most representative b-value when derived from images with different b-values. Required if Frame Type (0008,9007) Value 1 of this frame is ORIGINAL. May be present otherwise.
>Diffusion Directionality	(0018,9075)	1C	Specifies whether diffusion conditions for the frame are directional, or isotropic with respect to direction.

Publications and Presentations

- Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology Standards for Quantitative Imaging Biomarkers. *Radiology*. 2015 Aug 12. Epub ahead of print. doi: 10.1148/radiol.2015142202.
- Kessler, LG, et. al., The Emerging Science of Quantitative Imaging Biomarkers Terminology and Definitions for Scientific Studies and Regulatory Submissions, *Stat Methods Med Res 0962280214537333, first published on June 11, 2014 as doi:10.1177/0962280214537333*
- Raunig, DL, et. al., Quantitative Imaging Biomarkers: A Review of Statistical Methods for Technical Performance Assessment, *Stat Methods Med Res 0962280214537344, first published on June 11, 2014 as doi:10.1177/0962280214537344*
- Obuchowski, NA, et. al., Quantitative Imaging Biomarkers: A Review of Statistical Methods for Computer Algorithm Comparisons, *Stat Methods Med Res 0962280214537390, first published on June 11, 2014 as doi:10.1177/0962280214537390*
- Obuchowski, NA, et. al., Statistical Issues in the Comparison of Quantitative Imaging Biomarker Algorithms Using Pulmonary Nodule Volume as an Example, *Stat Methods Med Res 0962280214537392, first published on June 11, 2014 as doi:10.1177/0962280214537392*
- Huang, EP, et. al., Meta-analysis of the Technical Performance of an Imaging Procedure: Guidelines and Statistical Methodology, *Stat Methods Med Res 0962280214537394, first published on May 28, 2014 as doi:10.1177/0962280214537394*
- L. A. Pierce, B. F. Elston, D. A. Clunie, D. Nelson, and P. E. Kinahan, A Digital Reference Object to Analyze Calculation Accuracy of PET Standardized Uptake Value, *Radiology*, p. 141262, May 2015, *first published as 10.1148/radiol.2015141262*

QIBA Projects and activities have been funded in whole or in part with Federal funds from the National Institute of Biomedical Imaging and Bioengineering, National Institutes of Health, Department of Health and Human Services, under Contract Nos. **HHSN268201000050C** and **HHSN268201300071C.**