Variations in DCE-MRI Assessment of Breast Cancer Therapy Response: A Multicenter Data Analysis Challenge

Wei Huang[1], Xin Li[1], Xia Li[2], Ming-Ching Chang[3], Matthew J Oborski[4], Dariya I Malyarenko[5], Mark Muzi[6], Guido H Jajamovich[7], Andriy Fedorov[8], Yiyi Chen[1], Alina Tudorica[1], Sandeep N Gupta[3], Charles M Laymon[4], Kenneth I Marro[6], Hadrien A Dyvorne[7], James V Miller[3], Thomas L Chenevert[5], Thomas E Yankeelov[2], James M Mountz[4], Paul E Kinahan[6], Ron Kikinis[8], Bachir Taouli[7], Fiona Fennessy[8], and Jayashree Kalpathy-Cramer[9]

[1]Oregon Health & Science University, Portland, Oregon, United States, [2]Vanderbilt University, Nashville, Tennessee, United States, [3]General Electric Global Research, Niskayuna, New York, United States, [4]University of Pittsburgh, Pittsburgh, Pennsylvania, United States, [5]University of Michigan, Ann Arbor, Michigan, United States, [6]University of Washington, Seattle, Washington, United States, [7]Icahn School of Medicine at Mount Sinai, New York, New York, United States, [8]Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States, [9]Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, United States

Introduction: As a noninvasive imaging method for tissue microvasculature characterization, dynamic contrast-enhanced MRI (DCE-MRI) is increasingly being used in research settings and early phase clinical trials to assess tumor therapeutic response (1,2). Pharmacokinetic (PK) analysis of DCE-MRI data allows quantification of tissue biological properties *in vivo*, such as K^{trans} and v_e. However, the accuracy and precision of these parameters can be affected by a plethora of factors, including errors in determinations of pre-contrast T_1 (T_{10}) (3) and arterial input function (AIF) (2), as well as model selection for data fitting (4). A recent study (5) comparing four commercial software packages showed considerable variability in PK parameter quantifications even though all four tools are presumably based on the same Tofts model (6). Despite significant efforts in generating guidelines (7,8) for quantitative DCE-MRI studies, thorough comparison and validation of data analysis algorithms are needed before quantitative DCE-MRI is ready to be integrated into clinical practice. In this multicenter study, we sought to evaluate variations in DCE-MRI assessment of breast cancer response to therapy as a result of differences in PK models and software tools used.

Methods: A total of 7 institutions from the National Cancer Institute-sponsored Quantitative Imaging Network (QIN) participated in this DCE-MRI data analysis challenge, in which each center processed the shared DCE-MRI data with site-specific employment of PK models and software tools. The shared Breast DCE-MRI data were collected at one center from women with locally advanced breast cancer undergoing neoadjuvant chemotherapy (NACT), at multiple time points before, during, and after NACT. Axial bilateral DCE-MRI images were acquired with a 3D GRE-based TWIST sequence (9) using a 3T Siemens scanner. 32-34 image frames of 120-128 image slices each were acquired for ~ 10 min with 18-20 s temporal resolution.

The visit 1 (V1, before NACT) and V2 (after one NACT cycle) DCE-MRI data from 10 patients (for a total of 20 data sets) were shared anonymously for multicenter data analysis. In all, a total of 12 algorithms/software packages were used, including six based on the Tofts model (TM), four on the extended TM (ETM) (10), and two on the Shutter-Speed model (SSM) (11). These tools were built in-house, obtained as free open source, purchased commercially, or developed as prototype research software by a major manufacturer. To minimize variations in derived PK parameters caused by factors other than differences in models and algorithms, tumor ROI drawings (on multiple slices), a population-averaged AIF (9), and the measured mean tumor ROI T_{10} value were provided for data analysis by the center where the data were acquired. Mean tumor K^{trans}, v_e, k_{ep} (= K^{trans}/v_e), v_p (ETM only), and τ_i (mean intracellular water lifetime, SSM only) values for each patient at each visit, as well as their % changes (V2 relative to V1), were reported by each center. To assess reproducibility of each DCE-MRI parameter across all 12 algorithms and within each PK model, linear mixed models were fitted to obtain estimates of intra-class correlation coefficient (ICC) and within-subject coefficient of variation (wCV), and the corresponding 95% confidence intervals (CIs). Correlating with pathologic complete response (pCR) and non-pCR status, univariate logistic regression (ULR) models were fitted to evaluate capabilities of these parameters for early prediction of therapy response.

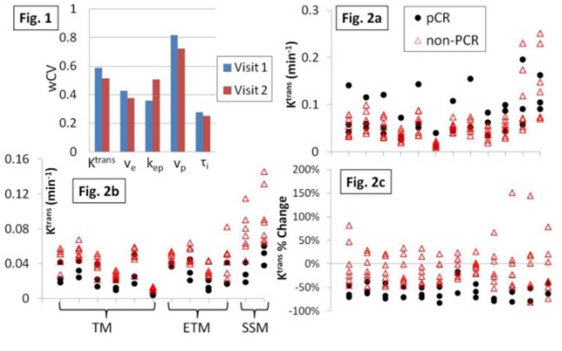
Table. K^{trans} Intra-Class Correlation Coefficient (ICC)

Algorithm Comparison	V1	V2	% Change (V21)
All algorithms (n = 12)	0.27 (0.11, 0.53)	0.18 (0.06, 0.44)	0.69 (0.46, 0.85)
TM (n = 6)	0.44 (0.20, 0.71)	0.17 (0.04, 0.54)	0.91 (0.79, 0.96)
ETM (n = 4)	0.62 (0.34, 0.84)	0.36 (0.11, 0.71)	0.63 (0.34, 0.84)
SSM (n = 2)	0.84 (0.58, 0.95)	0.72 (0.36, 0.92)	0.89 (0.69, 0.97)

Results: The column graph in **Fig. 1** shows the mean wCV values of the five DCE-MRI parameters across all 12 algorithms at V1 and V2, with K^{trans} wCV as high as 0.59 and v_p having the highest wCV of 0.82. wCV values were all reduced when only the algorithms within the same PK model were compared. The K^{trans} wCV was as low as 0.21 for the two SSM algorithms, similar to that of τ_i (Fig. 1) which is a SSM-only parameter. The **Table** lists K^{trans} ICC values and 95% CIs across all algorithms, and within TM, ETM, and SSM, respectively. ICC represents the proportion of total variation contributed by between-subject difference, with high ICC indicating good agreement among different measurement approaches. The K^{trans} ICC was generally increased from comparing all algorithms to comparing algorithms within one model. It is also important to note that K^{trans} % change generally had better agreement among the algorithms than absolute K^{trans} value at either V1 or V2. Similar ICC patterns were observed for the other DCE-MRI parameters.

Pathological analyses of post-NACT resection and pre-NACT biopsy specimens revealed 3 pCRs and 7 non-pCRs among the 10 patients. **Fig. 2** shows scatter plots of V1 (2a) and V2 (2b) K^{trans} of the pCRs (black) and non-pCRs (red) and the corresponding % changes (2c), with each column representing results returned from one algorithm. The columns associated with the same PK model were grouped next to each other in Figs. 2a-2c (labeled in Fig. 2b). For any of the 12 algorithms, V1 K^{trans} was not a good early predictor of response. However, almost all algorithms achieved good to excellent early discrimination of pCR and non-pCR using K^{trans} and k_{ep} (results not shown here) parameters at V2 and their % changes, with majority of them having ULR c statistics values equal to 1 (indicating complete separation of pCR and non-pCR) or greater than 0.9.

Discussion and Conclusion: This multicenter DCE-MRI data analysis challenge focuses on evaluation of variations in DCE-MRI prediction of breast cancer therapy response caused by differences in PK models and associated software algorithms. Despite the fixed inputs of tumor ROI definition, AIF, and T_{10} for analysis of the shared data sets with each algorithm, there are considerable variations in returned DCE-MRI parameters. This can be partly explained by the involvement of two SSM algorithms, which returned significantly ($P < 0.05$, analysis not shown here) greater K^{trans} (Figs. 2a and 2b) and v_e values than either TM or ETM. Parameter agreement is generally improved when comparing algorithms within the same model (Table). The relatively low temporal resolution of the raw data likely led to low precision and high variance in ETM estimation of the v_p parameter (4). Other contributions to parameter variations may come from differences in fixed physiological parameters, scaling factors, and goodness of fitting criteria that are employed in each algorithm. These factors were not controlled in this data analysis challenge. It is very encouraging within the context of therapy response assessment, however, that V2 K^{trans} (or k_{ep}) and its % change obtained from all 12 algorithms provided good to excellent early prediction of response. The predictive ability of the V2 parameters may be simply because the substantial decreases in pCR tumor perfusion and permeability outweighed parameter variations in this particular study setting, while that of the parameter % change is probably due to cancellation of the inter-algorithm systematic errors (or variations) in % change calculation, suggesting that parameter % change should be the imaging metric of choice in multicenter DCE-MRI trials employing different data analysis software tools.



Grant Support: NIH U01CA154602, U01CA142565, U01CA148131, U01CA166104, U01CA172320, U01CA151261, U01CA140230, U01CA154601, U54EB005149

References: **1.** O'Connor *et al. Nat Rev Clin Oncol* 2012;9:167-177. **2.** Yankeelov *et al. Curr Med Imaging Rev* 2009;3:91-107. **3.** Schabel *et al. Phys Med Biol* 2008;53:2345-2373. **4.** Sourbron *et al. Phys Med Biol* 2012;57:R1-R33. **5.** Heye *et al. Radiology* 2013;266:801-811. **6.** Tofts *et al. Magn Reson Med* 1991;17:357-367. **7.** Leach *et al. Eur Radiol* 2012;22:1451-1464. **8.** Evelhoch *et al. Cancer Res* 2005;65:7041-7044. **9.** Tudorica *et al. Magn Reson Imaging* 2012;30:1257-1267. **10.** Tofts *et al. J Magn Reson Imaging* 1999;10:223-232. **11.** Yankeelov *et al. Magn Reson Med* 2003;50:1151-1169.