SCIENTIFIC REPORTS

Correction: Publisher Correction

OPEN

Received: 5 May 2017
Accepted: 18 August 2017
Published online: 11 September 2017

# A Multi-Institutional Comparison of Dynamic Contrast-Enhanced Magnetic Resonance Imaging Parameter Calculations

**Joint Head and Neck Radiotherapy-MRI Development Cooperative***

**Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) provides quantitative metrics (e.g. $K^{trans}$, $v_e$) via pharmacokinetic models. We tested inter-algorithm variability in these quantitative metrics with 11 published DCE-MRI algorithms, all implementing Tofts-Kermode or extended Tofts pharmacokinetic models. Digital reference objects (DROs) with known $K^{trans}$ and $v_e$ values were used to assess performance at varying noise levels. Additionally, DCE-MRI data from 15 head and neck squamous cell carcinoma patients over 3 time-points during chemoradiotherapy were used to ascertain $K^{trans}$ and $v_e$ kinetic trends across algorithms. Algorithms performed well (less than 3% average error) when no noise was present in the DRO. With noise, 87% of $K^{trans}$ and 84% of $v_e$ algorithm-DRO combinations were generally in the correct order. Low Krippendorff's alpha values showed that algorithms could not consistently classify patients as above or below the median for a given algorithm at each time point or for differences in values between time points. A majority of the algorithms produced a significant Spearman correlation in $v_e$ of the primary gross tumor volume with time. Algorithmic differences in $K^{trans}$ and $v_e$ values over time indicate limitations in combining/comparing data from distinct DCE-MRI model implementations. Careful cross-algorithm quality-assurance must be utilized as DCE-MRI results may not be interpretable using differing software.**

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer worldwide[1]. Its 5-year survival rate has failed to improve from about 60% despite advances in imaging, surgery, radiotherapy targeting, and chemotherapy[2]. Thus, researchers are striving to individualize therapy for HNSCC to improve survival rates while limiting toxic effects in normal tissue, such as xerostomia, which can impact a patient's quality of life. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a noninvasive tool for examination of the microvasculature of tumors and normal tissue. The perfusion and permeability metrics estimated from pharmacokinetic modeling of DCE-MRI data may provide an indirect measure of tumor hypoxia, a condition associated with poor prognosis in HNSCC[3,4]. Therefore, it may be possible to build prognostic models to help tailor HNSCC treatments to individual patients based on that patient's DCE-MRI signature.

Investigators have used DCE-MRI to assess therapeutic response of HNSCC and have shown associations between DCE-MRI metrics and changes in salivary glands and mandible[5–11]. To the best of our knowledge, its use as a prognostic tool to inform treatment decisions for HNSCC has yet to be investigated in a large multisite prospective trial. Before such trials can begin, DCE-MRI inter-algorithm comparisons must be conducted to ensure consistency of output parameter maps for collating data during the multi-institution trial. Two quantitative metrics for DCE-MRI are the transfer constant for contrast agent transport from the blood plasma into the extravascular extracellular space ($K^{trans}$) and the volume fraction of the extravascular extracellular space ($v_e$). The calculation of these quantitative metrics can be impacted by the acquisition parameters. The accuracy and precision of these quantitative metrics can be influenced by arterial input function (AIF) quantification, temporal resolution in data acquisition, signal-to-noise ratio (SNR), and pharmacokinetic model selection[12–22]. For example, uncertainties in T1 map values and applied flip angle have been reported to cause errors of 88% in $K^{trans}$ and 73% in $v_e$, while reduced temporal resolution by 7-fold have reported decreses in $K^{trans}$ up to 48%[19]. Therefore,

*A comprehensive list of consortium members appears at the end of the paper. Correspondence and requests for materials should be addressed to C.D.F. (email: cdfuller@mdanderson.org)

acquisition parameters must be thoroughly tested and uniform across patients as they can dramatically impact measured DCE-MRI parameters.

The Tofts-Kermode pharmacokinetic model[23] is the most commonly used model for DCE-MRI analysis, but implementation of each algorithm differs in facets such as data preprocessing, approaches to numerical optimizations in kinetic analysis, and data postprocessing, which may impact the values of the output quantitative metrics. Several recent studies demonstrated significant inter-algorithm variability when evaluating DCE-MRI of the female pelvis, breast, and rectum[24–26]. Of these studies, the one by Huang *et al.*[25] demonstrated systematic differences in output parameter values between algorithms, which meant that results from different algorithms could be used together if correction factors were applied; the other studies, however, did not demonstrate any systematic errors. In addition, Cron *et al.*[27] found that the percentage of nonphysical values (e.g. $v_e$ values greater than 1) in the quantitative metrics increased as noise increased when they tested using three software packages. This noise dependence and inter-algorithm variance in quantitative DCE-MRI metrics are large obstacles to the clinical implementation of DCE-MRI and must be thoroughly investigated before proceeding with large multisite clinical trials using DCE-MRI in HNSCC patients.

In this study, we investigated the variability in $K^{trans}$ and $v_e$ across algorithms that are based on the Tofts-Kermode and extended Tofts pharmacokinetic models[28,29]. For this purpose, we used digital reference objects (DROs) from the Radiological Society of North America Quantitative Imaging Biomarkers Alliance[30] and DCE-MRI data from oropharyngeal squamous cell carcinoma patients who underwent multiple DCE-MRI scans during treatment with definitive chemoradiotherapy.

## Results

**DROs.** One of the Tofts-Kermode algorithms (algorithm 11) could not process the DROs because of the algorithm's structure. Therefore, the remaining 10 algorithms were used for DRO analysis. For the noiseless DRO, the stratified permutation test demonstrated that both $K^{trans}$ and $v_e$ were statistically significantly ordered correctly ($p < 0.05$) for all of the algorithms. Eighty-two percent of pairwise algorithm comparisons were statistically significantly different ($p < 0.05$) regarding $K^{trans}$, and 69% of the comparisons were statistically significantly different ($p < 0.05$) regarding $v_e$ based on the Wilcoxon rank-sum test. Figure 1 shows the algorithm performance for the noiseless DRO. Most of the $K^{trans}$ and $v_e$ measured values in the noiseless DRO were close to the true simulated values: 96% of $K^{trans}$ and 96% of $v_e$ measured values were within 10% of the simulated values. More spread in the measured values was observed at higher simulated values of $K^{trans}$ or $v_e$. Heat maps of the percentage error of $K^{trans}$ and $v_e$ measured values in comparison to the simulated values are shown in the supplemental material (Supplemental Fig. 1).

The stratified permutation test for the 28 DROs with noise demonstrated that in 86% and 84% of the cases (algorithm-DRO combinations), $K^{trans}$ and $v_e$ were statistically ordered correctly ($p < 0.05$) when one of the algorithms was excluded because of missing $K^{trans}$ values and failure of the $v_e$ test for all 28 of these DROs. Most of the test failures occurred at the lowest SNR (0.18). Eighty-four percent of the $K^{trans}$ pairwise comparisons and 81% of the $v_e$ pairwise comparisons were statistically significantly different ($p < 0.05$) based on the Wilcoxon rank-sum test results.

Heat maps of the percent error in $K^{trans}$ and $v_e$ relative to the simulated values in the 28 DROs with noise are shown in Fig. 2. The maximum percent error in this figure was set to 100% and the minimum percent error was set to −100%. Therefore, any $K^{trans}$ and $v_e$ values greater than the maximum percent error are mapped to red. The only trend found was less error at higher $K^{trans}$ and $v_e$ simulated values although there is more spread in the measured values at these higher $K^{trans}$ and $v_e$ simulated values. Algorithms that used spatial averaging were found to have statistically significantly less ($p < 0.05$) $K^{trans}$ and $v_e$ calculated error than algorithms that did not have spatial averaging according to the student's t-tests.

We observed large variation in the percentage of values removed due to the threshold for $K^{trans}$ and $v_e$ for each algorithm. Some algorithms had almost no values removed, and some had a median of 70% of values removed.

These DRO results are for one method of excluding $K^{trans}$ and $v_e$ values. We also analyzed the data using the central 95% of the data for each $K^{trans}$-$v_e$ pair with no threshold restrictions, which produced consistent test results.

**Patients.** The percentages of $K^{trans}$ and $v_e$ values removed from patient ROIs because they were outside the bounds of the threshold are shown in Fig. 3 for the pretreatment, midtreatment, and posttreatment $K^{trans}$ and $v_e$. As in the DROs, the percentages varied: some algorithms had low percentages removed, implying that they mostly produced realistic values, whereas some algorithms produced almost nothing but unrealistic values for certain patients. The average percentage removed for $K^{trans}$ was 27%, 26%, and 22% for pretreatment, midtreatment, and posttreatment respectively. The average percentage removed for $v_e$ was 46%, 49%, and 48% for pretreatment, midtreatment, and posttreatment respectively.

According to results of the likelihood ratio test, all differences were statistically significantly ($p < 0.05$) dependent upon the algorithm except for the pretreatment-to-posttreatment change in $K^{trans}$ when all algorithms were included in the model. Algorithms were subset into Tofts-Kermode and extended Tofts groups. In the Tofts-Kermode group, three changes were not statistically significantly dependent on algorithm ($p < 0.05$): pretreatment-to-midtreatment change in $K^{trans}$, midtreatment-to-posttreatment change in $K^{trans}$, and midtreatment-to-posttreatment change in $v_e$. In the extended Tofts group, algorithm was not a significant factor ($p < 0.05$) in pretreatment-to-posttreatment change. In all other changes, the algorithm was a significant factor. In all linear mixed effects models, the variance explained by the ROI was much smaller than the residual variance, suggesting that the ROI does not explain much of the variation seen in the linear mixed effects model. All organ variance was less than 30% of the residual variance; on average, it was 8% of the residual variance.

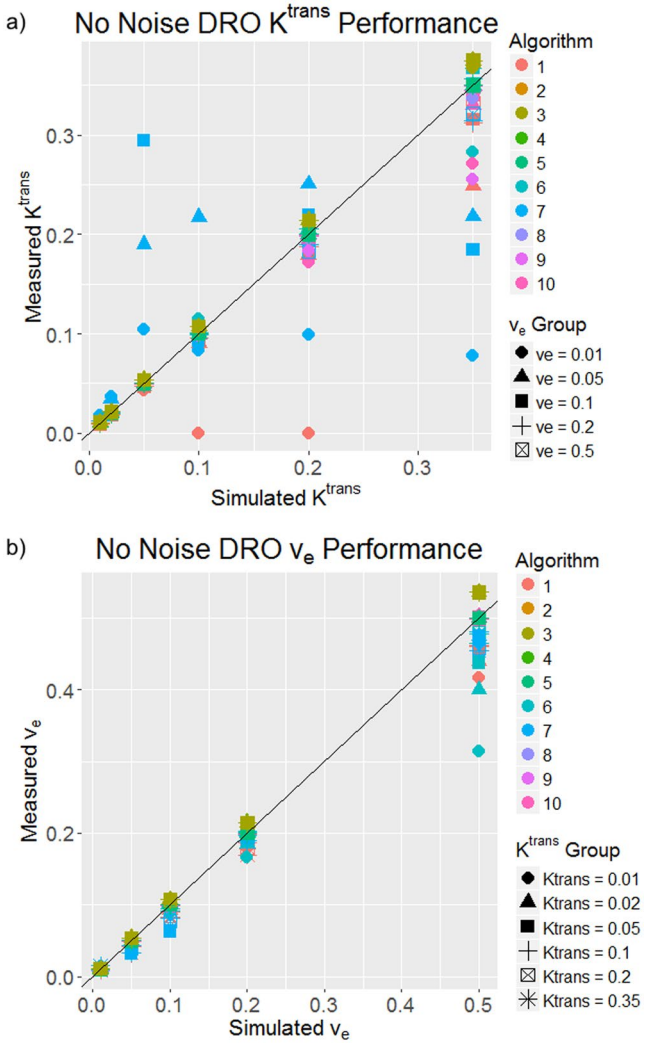


**Figure 1.** Plots of algorithm performance in a DRO with no noise for (**a**) $K^{trans}$ and (**b**) $v_e$. The simulated values are on the x-axis, and the measured values from each algorithm are on the y-axis. The 45° line represents 100% accuracy of the measured values. Each color represents a different algorithm, and each shape represents a different $v_e$ column in (**a**) and a different $K^{trans}$ row in (**b**).

Figure 4 demonstrates an example of the difference in parameter values exported from different algorithms. The $K^{trans}$ maps from the same axial slice of a patient are shown for all algorithms. It can be seen that some algorithms output mostly lower $K^{trans}$ values while others output mostly higher $K^{trans}$ values. In addition, some algorithms fit the noise data in voxels outside of the anatomy while other algorithms generated $K^{trans}$ maps only within the anatomy.

Carletta's thresholds for good agreement between algorithms ($\alpha \geq 0.8$) and sufficient agreement for tentative conclusions ($0.800 > \alpha > 0.667$) were used[31] to assess the results of Krippendorff's alpha tests. The tests were run using all of the algorithms and also subsets of the algorithms, which were placed into Tofts-Kermode and extended Tofts groups. Of all of these tests, only those in the extended Tofts group had alphas that fell in range for tentative conclusions: 7 of the 108 tested correlations in this group had alphas in this tentative conclusions range. No alphas were in the good agreement range. An illustration of this inconsistent sorting of patients is shown in the supplemental material (Supplemental Fig. 3). Carletta's thresholds for good agreement and tentative conclusions are weaker than those suggested by others. Krippendorff[32] and Neuendorf[33] suggested using higher standards, which would remove all the metrics found to be partially reliable across algorithms.

Few statistically significant Spearman correlations ($p < 0.05$) were observed: 8% of all tested $K^{trans}$ correlations and 29% of all tested $v_e$ correlations across all aglorithms. The only trend in these correlations across algorithms was a statistically significant Spearman correlation of $v_e$ in the GTV-P.

## Discussion

Use of DCE-MRI is increasing in oncology research and investigators have performed many promising studies indicating correlations between predicted therapeutic outcome and DCE-MRI metrics[7,8]. However, many different DCE-MRI platforms were employed in these studies, and no studies have demonstrated whether data and

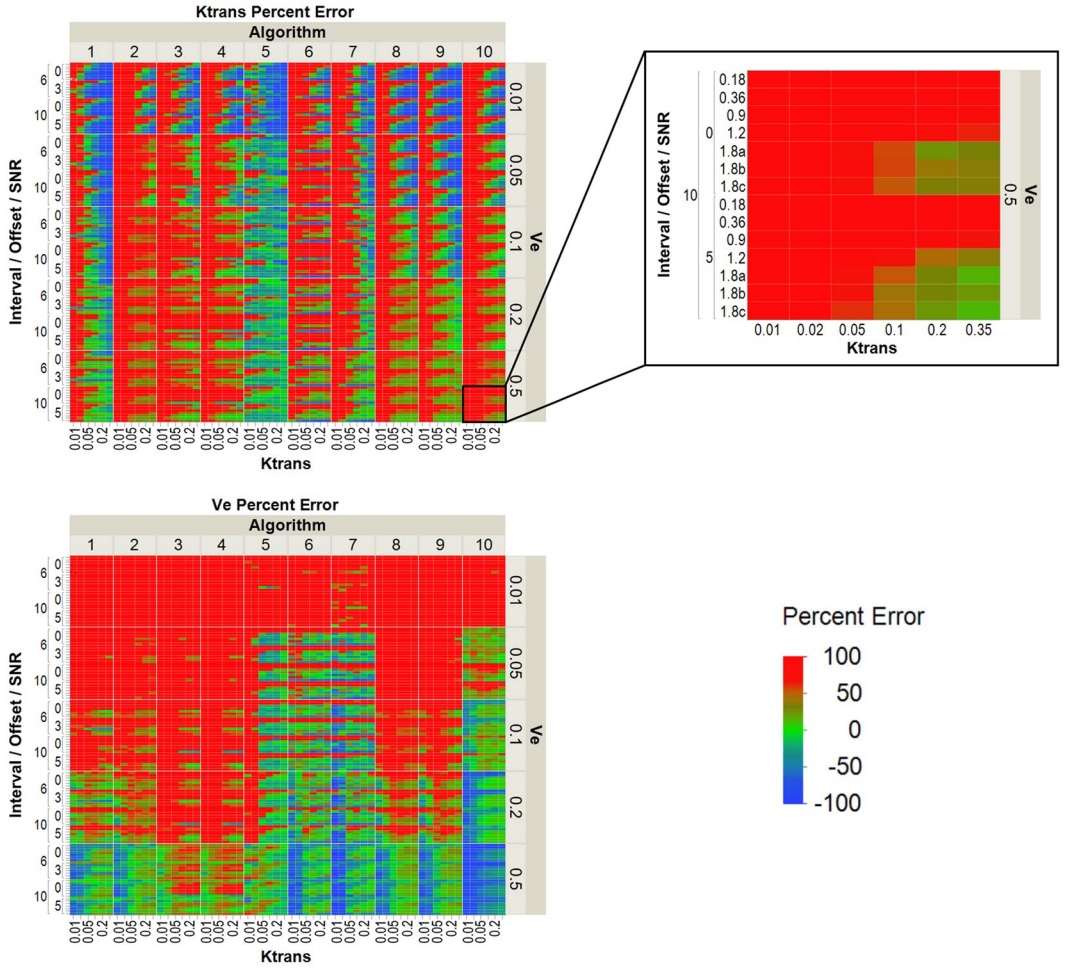


**Figure 2.** Heat maps of the percentage error for $K^{trans}$ (top left) and $v_e$ (bottom left) by algorithm in the 28 DROs with noise. The percentage error is defined using the formula ([measured − simulated]/simulated *100). The left side of the heat map is grouped by the timing interval used for the DRO (6 or 10 s), the timing offset used for the DRO (0 or 3 s for the 6 s timing interval, 0 or 5 s for the 10 s timing interval), and the SNR (0.18–1.8). The inset (top right) shows the $K^{trans}$ and SNR values for each block in the heat maps. The maximum percentage error is defined as 100%, and the minimum percentage error is set to −100%. Any errors greater than the maximum percentage are also mapped as 100% error in color. Each DRO is differentiated by its sampling interval, timing offset, and SNR as determined by the $S_0$ and sigma value used to create the DRO.

conclusions regarding HNSCC can be aggregated. We addressed this issue by analyzing the same sets of DRO and HNSCC patient data with a subset of the currently used algorithms that are based on the Tofts-Kermode or extended Tofts model, as these pharmacokinetic models are the ones most commonly employed in DCE-MRI.

The key results from this study are that algorithms were able to determine high values from low values on DROs, but workflow differences may obscure the ability to discern values across algorithms in patients. This may be specifically related to T1 mapping which was not controlled in the patient portion of this study. Specifically, trends among algorithms from the same institution (institution supplied both Tofts-Kermode and extended Tofts algorithms) were consistent, but not across institutions. This highlights the effect of preprocessing, also shown by the impact of spatial averaging on the calculated error. Therefore, translatability of DCE-MRI across algorithms is not currently feasible.

A digital phantom was used to assess algorithms with a known "ground truth". The DROs we used had SNRs of 0.18 to 1.80 in the noisy DROs. Although these SNR values and $K^{trans}$ and $v_e$ values within the DRO are below that typically found in head and neck cancer cases[34–37], the DROs were used due to their availability and Quantitative Imaging Biomarkers Alliance-backed quality. The DROs, however, do not come with instructions for interpretation of results, which makes conclusions difficult especially for the DROs that contain very high noise.

The good algorithm performance for the noiseless DRO is consistent with the results reported by Huang *et al.*[25] and suggest that the algorithms tested here are constructed properly. However, the error increased dramatically when high levels of noise were added to the images. Our assessment using percentage error may explain why the error appeared extremely high in the low $K^{trans}$ and $v_e$ regions as a small absolute error in this region will appear with a high percentage error. Heat maps of the error with the noisy DROs are shown in the supplementary data (Supplemental Fig. 2) to remove this discrepancy in percentage error between low and high values.

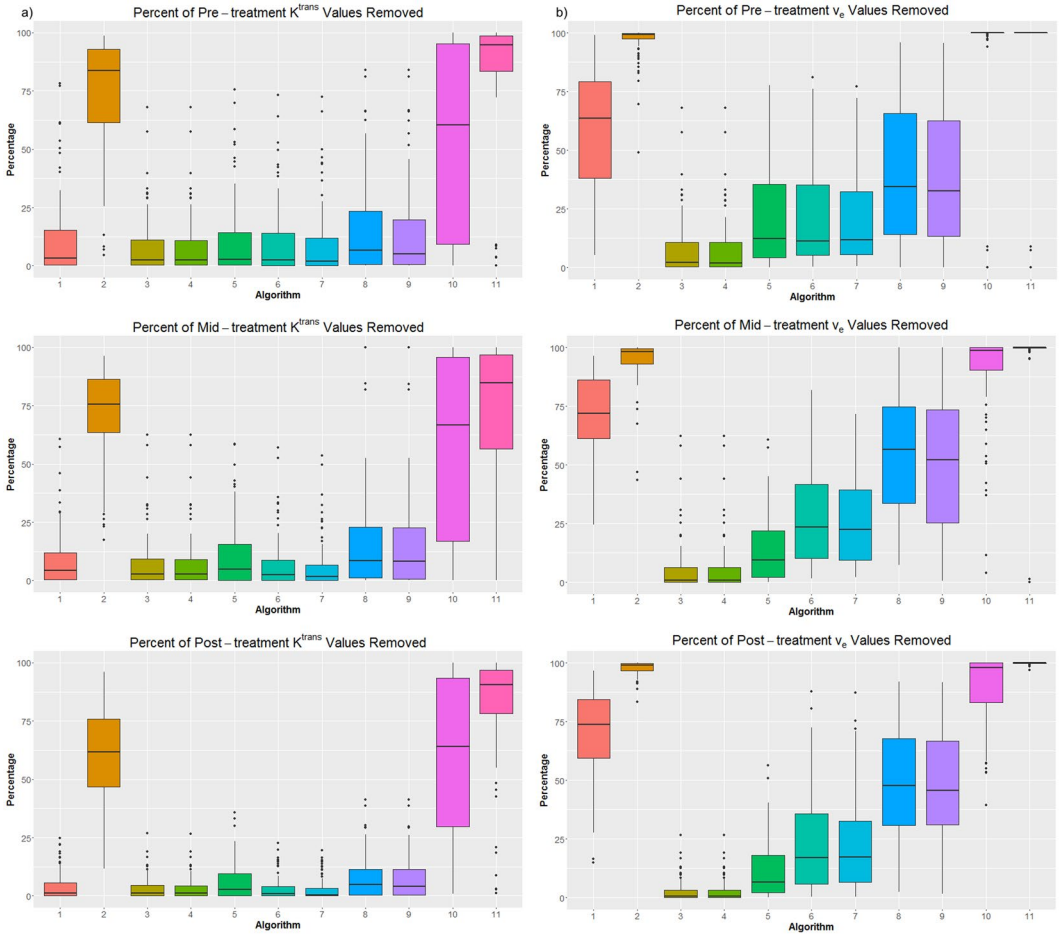


**Figure 3.** Percentages of (**a**) $K^{trans}$ and (**b**) $v_e$ values removed from patient images. The boxplots for each algorithm include the percentages removed for all patients and contours.

The difference between algorithms was significant for DROs according to the Wilcoxon rank-sum test results, which is consistent with the results reported by Beuzit *et al.*[26], who used SNRs of 10 and 100 and still found significant differences between different software packages. A limitation of this test is that if the differences between two algorithms are small but all of one sign (such that all values from one algorithm are higher than all values from another algorithm), the differences will be statistically significant. This does not appear to be the cause of the statistically significant differences observed here because each algorithm has its own error signature, and we could not identify a systematic error in any of the algorithms.

The DRO results demonstrated the potential of DCE-MRI quantitative metrics for clinical application, an illustration of which was the patient data set we used. The significance of including algorithm in the linear mixed effects models was consistent with the Wilcoxon rank-sum test results for the DROs. The small variance explained by the ROI compared with the residual variance in the linear mixed effects models was surprising. If associations can be found using DCE-MRI, different trends between normal tissue and tumor would particularly be expected, yet the ROI provided little explanation of the variance in the data in the linear mixed effects models.

A majority of the algorithms tested produced statistically significant Spearman correlations of $v_e$ in the GTV-P. The agreement of Spearman correlations across algorithms within the GTV-P but not within normal tissue may be due to a difference in contrast-induced signal change, as the GTV-P has a much higher signal change than does normal tissue in DCE-MRI. This means that the GTV-P has higher $K^{trans}$ and lower error in the presence of noise based on the DRO data. However, this agreement of Spearman correlations of $v_e$ in the GTV-P is contradicted by the Krippendorff's alpha results for the GTV-P. Only the midtreatment $K^{trans}$ value in the extended Tofts group had an alpha in the range where tentative conclusions can be drawn. This discrepancy may be explained by small interpatient variability in the $K^{trans}$ and $v_e$ values, which limited the algorithms' ability to separate patients into above or below the median. However, the Spearman rank correlation coefficient identifies trends and is not as affected by interpatient differences in values as Krippendorff's alpha if the trend is consistent.

The Krippendorff's alpha results demonstrated that different algorithms do not consistently classify patients' $K^{trans}$ and $v_e$ values, change in values, or percent change in values. These results indicate that there is currently no clinical level at which these quantitative metrics can be used across algorithms to quantify patients. Based on the algorithms' performance for the DROs in the stratified permutation test in our study, this result from Krippendorff's alpha tests is surprising. However, small interpatient variation in the $K^{trans}$ and $v_e$ values may have

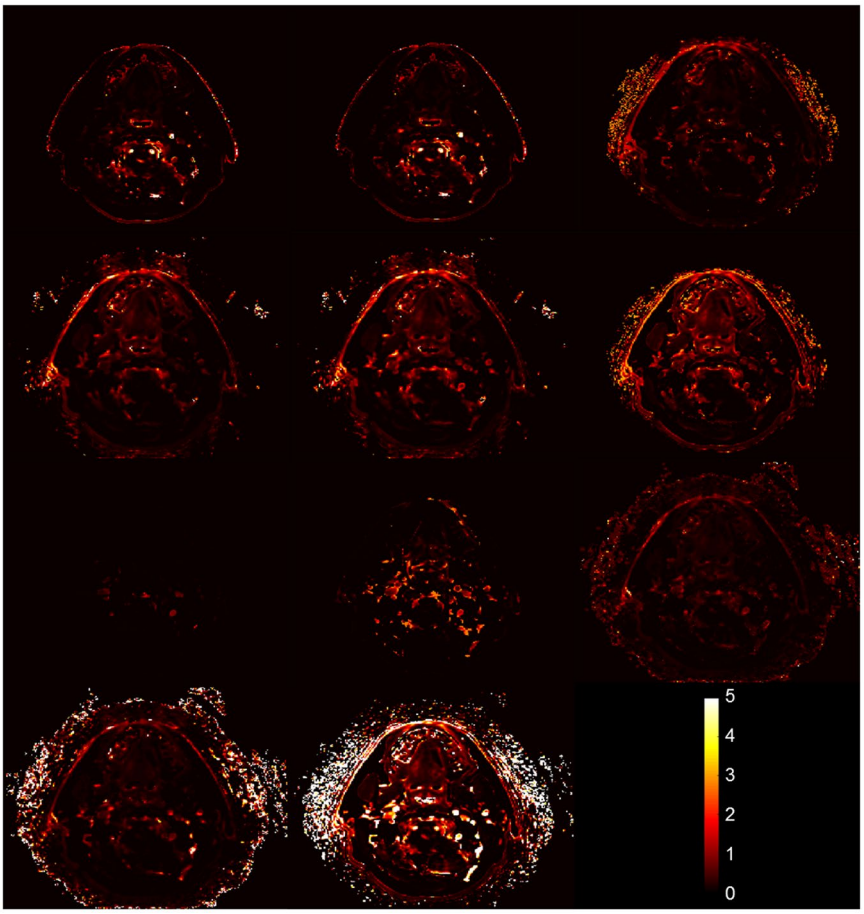


**Figure 4.** Illustration of differences in $K^{trans}$ ($min^{-1}$) maps exported by different algorithms for one axial DCE-MRI slice.

caused the low inter-algorithm reliability. This low inter-algorithm reliability, even within the Tofts-Kermode and extended Tofts groups, contrasts with the results described by Huang *et al.*[25]. They found good parameter agreement for percentage change when they grouped algorithms by pharmacokinetic model and that all of the algorithms provided good prediction of response to therapy as assessed using univariate logistic regression. This difference may have resulted from the imaging technique used, tissue of interest, and/or patient distribution of $K^{trans}$ and $v_e$ values.

Uncertainties in DCE-MRI exist due to AIF selection and imaging parameters[12,13,16–18,22], but we did not explore them in this study because they were controlled: we examined each algorithm with the same patient DCE-MRI images, variable-flip-angle images, and AIF. In previous studies, T1 mapping and AIF selection impacted $K^{trans}$ and $v_e$ values[12–17,22,38–40]. The agreement between algorithms that we observed may have been lower if we had included all of the differences typically seen in a multisite clinical trial, including different scanners, scanning protocols, AIFs, DCE-MRI algorithms at each institution. In our relatively controlled study, we observed statistically significant differences in both DRO and patient data among the algorithms. It must be acknowledged that there is no 'ground truth' against which these algorithms can be compared, and it is unclear whether there was a true therapeutic effect that should have been identified by DCE-MRI of patient data. Even if there was no net effect across this population of patients, however, it is clear that different approaches to DCE-MRI analysis have significant impact on within-patient trends.

We chose the upper bound for $K^{trans}$ since one of the algorithms in this study used 5 $min^{-1}$, providing a feasible physical upper limit. We chose the lower bound for $K^{trans}$ because when a given pixel or voxel has a poor fit within an algorithm, it is often given a value of 0 or a negative value. Accordingly, we excluded these values from analysis. We chose the bounds for $v_e$ based on the physical limits given by its definition as a fractional space. Furthermore, poor fits in an algorithm are often mapped to 0 or 1. Therefore, we excluded these values. While 0 is a physically realistic value for $K^{trans}$ and 0 and 1 are physically realistic values for $v_e$, these values must be excluded owing to a high proportion of bad pharmacokinetic model fits mapped to these values. The high percentage of values that must be removed represents an area of improvement for future algorithms. Cron *et al.*[27] demonstrated that as noise in DCE-MRI scans increases, the percentage of nonphysical $K^{trans}$ and $v_e$ values increases. Thus, voxel-based analysis of DCE-MRI quantitative metrics may not be reliable, so global metrics, such as average, of regions must be used for studies. For regions in which a high percentage of values are excluded, the average value extracted is not a reliable metric, as it comes from only a small subregion which is not representative of the whole region. This

| Institution | Model(s) Used |
|---|---|
| Massachusetts General Hospital | Tofts-Kermode (description in Supplemental Data) |
| MD Anderson Cancer Center | Tofts-Kermode and Extended Tofts (description in Supplemental Data) |
| Netherlands Cancer Institute | Tofts-Kermode and Extended Tofts[47] |
| nordicICE | Extended Tofts[48] |
| Oregon Health & Science University | Tofts-Kermode and Tofts-Kermode[14,49,50] |
| Princess Margaret Cancer Center | Tofts-Kermode and Extended Tofts[51,52] |
| University of Texas at Austin | Tofts-Kermode[53,54] |

**Table 1.** Description of Algorithms. Algorithms are listed in alphabetical order not order displayed in figures.

issue can be mitigated on the imaging end by increasing the SNR at the cost of the increased scan time, poorer temporal resolution, spatial resolution, or coverage, and potentially on the software end by improving how algorithms handle noise through the use of DROs.

In summary, we showed that rigorous standardization and careful quality assurance of software programs, including comparison of parameter calculations with standard data sets, are needed for collating pharmacokinetic analysis of DCE-MRI data among different algorithms. This must include assessment of the impact of image noise on quantitative metric error. Authors recently reported the need for careful quality assurance for functional MRI[41]. Efforts like those by the Quantitative Imaging Biomarkers Alliance to standardize DCE-MRI acquisition parameters represent a natural step forward for quality assurance and serve as the foundation for the current quality assurance work used in the present study.

To support these efforts, we provided our data set in a repository to allow for their use as perpetual head and neck cancer patient-derived standards for future DCE-MRI software and/or algorithm development in addition to the extant DRO library maintained by one of the authors (D. Barboriak). To that end, we recommend the following:

1. Consistent use of the same software for DCE-MRI analysis within a given study and for cross-comparisons between studies.
2. Specification and setting of acquisition parameters before proceeding with clinical trials as with the present data set.
3. Before performing multi-institution clinical trials, confirmation that DCE-MRI parameter values are consistent across institutions.
4. Inclusion of reference to a DRO with clinically relevant SNRs to benchmark performance of DCE-MRI software using clear evaluation criteria.

Clinically, our DRO data point to the fact that algorithms differed substantially despite reliance on the same basic underlying pharmacokinetic model(s), performing relatively stable in low-noise conditions. This, coupled with the inter-algorithm variability observed with the *in vivo* head and neck cases (which were performed in immobilization on a single MRI platform with standard AIF selection) suggests that, at present, any clinical trial desiring to implement DCE-MRI, should at a minimum, use a single pre-specified DCE-MRI software workflow, and eschew use of multiple algorithms. This also means that DCE-MRI findings from one software are broadly uninterpretable in a differing platform at present.

Until quantitative metrics can be reliably calculated across algorithms, patient-derived DCE-MRI analyses with different algorithms cannot be aggregated. Semiquantitative metrics, such as the area under the curve, have been shown to be more reproducible than quantitative metrics and may be the best interim option for use in prognostic studies using different algorithms[42]. Further refinement is required before DCE-MRI software-derived parameters can be used as a routine cross-institutional metric for multi-site clinical trials.

## Materials and Methods

**Algorithms.** Eleven algorithms from six institutions and one commercial software package were analyzed. They consisted of seven Tofts-Kermode models (identified herein as algorithms 2, 3, 5, 6, 8, 10, 11) and four extended Tofts models (algorithms 1, 4, 7, 9). Spatial averaging on the DCE-MRI images was used in algorithms 5, 6, 7, 8, and 9. All algorithms are currently used for research applications at the respective institutions. The algorithms are described in Table 1.

**DROs.** DROs provided by the Radiological Society of North America Quantitative Imaging Biomarkers Alliance were used to assess algorithm performance. The DROs had six $K^{trans}$ values ranging from 0.01 $min^{-1}$ to 0.35 $min^{-1}$ that were constant across the rows and five $v_e$ values ranging from 0.01 to 0.5 that were constant down the columns, resulting in 30 different $K^{trans}$-$v_e$ pairs, each encompassing $10 \times 10$ pixels. The $K^{trans}$ and $v_e$ values were used to generate synthetic image data using the Tofts-Kermode two-parameter model run in JSim, an open-source modeling system[23,43]. One DRO without noise[44] and 28 DROs with noise (SNR 0.18–1.8)[45] simulated by varying the sampling interval, timing offset, $S_0$, and sigma value were used to evaluate algorithm performance. For each $K^{trans}$-$v_e$ pair, the output pixels from the algorithms were subjected to a threshold to non-physiologic pixels ($0 < K^{trans}$ output $< 5$ and $0 < v_e$ output $< 1$) and then averaged.

| Patient Number | Sex | Age (years) | Race/ Ethnicity | Smoking Status | Primary Tumor Site | TNM Category | Chemotherapy (weekly) |
|---|---|---|---|---|---|---|---|
| 1 | M | 52 | White | N | Base of tongue | T3N1M0 | Cisplatin |
| 2 | M | 53 | White | Y | Base of tongue | T2N2aM0 | Cetuximab |
| 3 | M | 60 | White | Y | Tonsil | T4N2bM0 | Cisplatin |
| 4 | M | 55 | White | Y | Tonsil | T3N2bM0 | Cisplatin |
| 5 | M | 65 | White | N | Base of tongue | T2N1M0 | Cetuximab |
| 6 | M | 57 | Hispanic | Y | Tonsil | T2N2cM0 | Cisplatin |
| 7 | M | 60 | White | Y | Base of tongue | T2N2bM0 | Cisplatin |
| 8 | M | 58 | Black | Y | Base of tongue | T2N2cM0 | Cisplatin |
| 9 | M | 62 | Asian | Y | Tonsil | T4N2cM0 | Cisplatin |
| 10 | F | 48 | White | Y | Tonsil | T4N2bM0 | Cisplatin |
| 11 | M | 56 | White | N | Tonsil | T2N2cM0 | Cisplatin |
| 12 | M | 68 | White | Y | Tonsil | TxN2cM0 | Cisplatin |
| 13 | M | 47 | White | N | Tonsil | T3N2bM0 | Cisplatin |
| 14 | M | 47 | White | Y | Tonsil | T3N2bM0 | Cisplatin |
| 15 | M | 55 | White | N | Base of tongue | T4N2bM0 | Cisplatin |

**Table 2.** Study Patient Demographics.

**Patients.** Fifteen patients diagnosed with human papillomavirus-positive oropharyngeal squamous cell carcinoma were included in this study under a protocol approved by the institutional review board at MD Anderson Cancer Center. All patients gave their study-specific informed consent. All methods were performed in accordance with the relevant guidelines and regulations. Patients underwent DCE-MRI scans from December 2013 to October 2014. The criteria for study inclusion were an age older than 18 years, histologically documented stage III or IV human papillomavirus-positive oropharyngeal squamous cell carcinoma according to the American Joint Committee on Cancer 7$^{th}$ edition staging criteria, eligibility for definitive chemoradiotherapy, and an Eastern Cooperative Oncology Group performance status of 0 to 2. Patients were excluded for any of the following reasons: definitive resection of a primary tumor, administration of induction chemotherapy before radiotherapy, a prior cancer diagnosis except that of appropriately treated localized epithelial skin cancer or cervical cancer, prior radiotherapy to the head and neck, contraindications for gadolinium-based contrast agents, and claustrophobia.

Patient median age was 56 years (range, 47–68), with 14 men and 1 woman. All patients received radiotherapy at 70 Gy in 33 fractions. The majority of the patients (87%) received cisplatin-based chemotherapy concurrently with radiotherapy. Patient, disease, and treatment characteristics are listed in Table 2. Patient 12 did not have a primary tumor because he underwent bilateral tonsillectomy before scanning.

All patients underwent DCE-MRI scans within 1 week prior to treatment, 3–4 weeks after the start of treatment, and 6–8 weeks after the completion of treatment. The DCE-MRI scans were done using a 3.0 T Discovery 750 MRI scanner (GE Healthcare) with six-element flex coils and a flat insert table (GE Healthcare). The same immobilization devices (individualized head and shoulder mask, customized head support, and intraoral tongue-immobilizing/swallow-suppressing dental stent) were employed in longitudinal scans to improve image co-registration and to reduce interval physiologic motion (e.g., swallowing).

Thirty axial slices with a field of view of 25.6 cm and thickness of 4 mm were selected to cover the spatial region encompassing the palatine process region cranially to the cricoid cartilage caudally for all scans. Prior to DCE-MRI, T1 mapping was performed using a total of six variable-flip-angle three-dimensional spoiled gradient recalled echo sequences (flip angles: 2°, 5°, 10°, 15°, 20°, and 25°; repetition time/echo time, 5.5/2.1 ms; number of effective excitations, 0.7; spatial resolution, 2 mm × 2 mm × 4 mm; scan time, 3 minutes). The DCE-MRI acquisition consisted of a three-dimensional fast spoiled gradient recalled echo sequence to gain sufficient SNR, contrast, and temporal resolution. The following scan parameters were used: flip angle, 15°; repetition time/echo time, 3.6/1.0 ms; number of effective excitations, 0.7; spatial resolution, 2 mm × 2 mm × 4 mm; temporal resolution, 5.5 s; number of temporal frames, 56; pixel bandwidth, 326 Hz; acceleration factor, 2; and scan time, 6 minutes. Gadopentetate dimeglumine (Magnevist; Bayer HealthCare Pharmaceuticals) was administered intravenously to the patients at the end of the sixth frame (dose, 0.1 mmol/kg at a rate of 3 mL/second) followed by a 20-mL saline flush via a power injector (Spectris MR Injector; Medrad) at a rate of 3 mL/second.

Variable-flip-angle images, DCE-MRI images, and a bootstrapped population AIF measured in a region of interest in the carotid artery[11] were distributed to each institution to use in their algorithm(s) to generate $K^{trans}$ and $v_e$ parameter maps for each patient.

Each patient had 6 regions of interest (ROIs)—contralateral and ipsilateral parotid glands, contralateral and ipsilateral submandibular glands, sublingual glands, and a primary gross tumor volume (GTV-P)—contoured on his or her pretreatment images by a radiation oncologist with 7 years of experience (A.S.R. Mohamed). Midtreatment and posttreatment images were deformably registered to the pretreatment images using a

commercially available software program (Velocity AI, version 3.0.1; Varian Medical Systems). The deformation vector fields were exported from the deformation software and used with an in-house MATLAB code (MATLAB 2014b; MathWorks) to deform the ROIs and extract $K^{trans}$ and $v_e$ values from the six ROIs on each parameter map at the three time points. For each ROI, $K^{trans}$ and $v_e$ values were subjected to the same threshold constraints as in the DROs and then averaged.

**Statistical methods.** A stratified permutation test was designed to determine whether the $K^{trans}$ and $v_e$ values from an algorithm for a specific DRO were generally ordered correctly in the DRO. Permutation tests work by rearranging data in many possible ways in order to estimate the sampling distribution for the test statistic. Algorithms were compared on a pairwise basis using a paired Wilcoxon rank-sum test to determine if the outputs of two algorithms were statistically different (R software package, version 3.3.1). Algorithms were split into two groups based on if spatial averaging was used on the DCE-MRI scans. The two groups were compared using a one-sided student's t-test to determine if lower error on the DROs was calculated when spatial averaging was used. All p-values were adjusted using the Benjamini-Hochberg correction for multiple comparisons.

For patient DCE-MRI data, consistency of trends across algorithms was assessed using linear mixed effects models (R lme4 package, version 1.1.12) constructed for the differences between the pretreatment and midtreatment, pretreatment and posttreatment, and midtreatment and posttreatment quantitative metrics, and percent change in these three time differences. Two mixed effects models were created: one in which the algorithm was a fixed effect and the ROI was a random effect ($\Delta \sim \text{algorithm} + (1|\text{ROI})$) and one in which only the random effect of the ROI was included ($\Delta \sim 1 + (1|\text{ROI})$). A likelihood ratio test was performed for these two models to determine if the algorithm was a significant factor in the measured changes. All p-values were adjusted using the Benjamini-Hochberg correction for multiple comparisons (R, version 3.3.1). We used linear mixed effects models with likelihood ratio tests instead of ANOVA tests because in most comparisons we observed statistically different variances as determined using the Levene test, which violates one of the assumptions of ANOVA tests. Intraclass correlation coefficient is more appropriate for complete data sets[46], so it was not applicable for this data set.

For all ROIs, patients were categorized as above or below the median values from a given algorithm using three different metrics: (1) each time point, (2) difference between time points, and (3) percent difference between time points. Krippendorff's alpha was used to assess inter-algorithm reliability (R, irr package, version 0.84). We used Krippendorff's alpha to compare algorithms because of its ability to handle missing data, which occurred because for some algorithms, all $K^{trans}$ and $v_e$ values were outside the threshold for a given patient's ROI.

Trends within each algorithm were assessed using Spearman's rank correlation coefficient (R, version 3.3.1). Spearman correlations were conducted using three different sets of time points: (1) all three time points, (2) only the pretreatment and midtreatment time points, and (3) only the pretreatment and posttreatment time points were evaluated. All p-values were adjusted using the Benjamini-Hochberg correction for multiple comparisons. For all statistical tests, p-values below 0.05 after adjustment were considered significant.

## References


1. Jemal, A. *et al.* Global cancer statistics. *CA: a cancer journal for clinicians* **61**, 69–90, https://doi.org/10.3322/caac.20107 (2011).
2. Howlader, N. *et al.* SEER Cancer Statistics Review, 1975–2014, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975–2014/, based on November 2016 SEER data submission, posted to the SEER web site, April 2017.
3. Bernstein, J. M., Bernstein, C. R., West, C. M. & Homer, J. J. Molecular and cellular processes underlying the hallmarks of head and neck cancer. *European archives of oto-rhino-laryngology* **270**, 2585–2593, https://doi.org/10.1007/s00405-012-2323-x (2013).
4. Horsman, M. R., Mortensen, L. S., Petersen, J. B., Busk, M. & Overgaard, J. Imaging hypoxia to improve radiotherapy outcome. *Nature reviews clinical oncology* **9**, 674–687, https://doi.org/10.1038/nrclinonc.2012.171 (2012).
5. Houweling, A. C. *et al.* MRI to quantify early radiation-induced changes in the salivary glands. *Radiotherapy and oncology* **100**, 386–389, https://doi.org/10.1016/j.radonc.2011.08.020 (2011).
6. Juan, C. J. *et al.* Perfusion characteristics of late radiation injury of parotid glands: quantitative evaluation with dynamic contrast-enhanced MRI. *European radiology* **19**, 94–102, https://doi.org/10.1007/s00330-008-1104-9 (2009).
7. Bernstein, J. M., Homer, J. J. & West, C. M. Dynamic contrast-enhanced magnetic resonance imaging biomarkers in head and neck cancer: potential to guide treatment? A systematic review. *Oral oncology* **50**, 963–970, https://doi.org/10.1016/j.oraloncology.2014.07.011 (2014).
8. Noij, D. P. *et al.* Contrast-enhanced perfusion magnetic resonance imaging for head and neck squamous cell carcinoma: a systematic review. *Oral oncology* **51**, 124–138, https://doi.org/10.1016/j.oraloncology.2014.10.016 (2015).
9. Cheng, C. C. *et al.* Parotid perfusion in nasopharyngeal carcinoma patients in early-to-intermediate stage after low-dose intensity-modulated radiotherapy: evaluated by fat-saturated dynamic contrast-enhanced magnetic resonance imaging. *Magnetic resonance imaging* **31**, 1278–1284, https://doi.org/10.1016/j.mri.2013.03.018 (2013).
10. Lee, F. K., King, A. D., Kam, M. K., Ma, B. B. & Yeung, D. K. Radiation injury of the parotid glands during treatment for head and neck cancer: assessment using dynamic contrast-enhanced MR imaging. *Radiation research* **175**, 291–296, https://doi.org/10.1667/RR2370.1 (2011).
11. Cooperative, J. Ha. N. R.-M. D. Dynamic contrast-enhanced MRI detects acute radiotherapy-induced alterations in mandibular microvasculature: prospective assessment of imaging biomarkers of normal tissue injury. *Scientific reports* **6**, 29864, https://doi.org/10.1038/srep29864 (2016).
12. Huang, W. *et al.* The Impact of Arterial Input Function Determination Variations on Prostate Dynamic Contrast-Enhanced Magnetic Resonance Imaging Pharmacokinetic Modeling: A Multicenter Data Analysis Challenge. *Tomography* **2**, 56–66, https://doi.org/10.18383/j.tom.2015.00184 (2016).
13. Yankeelov, T. E. & Gore, J. C. Dynamic Contrast Enhanced Magnetic Resonance Imaging in Oncology: Theory, Data Acquisition, Analysis, and Examples. *Current medical imaging reviews* **3**, 91–107, https://doi.org/10.2174/157340507780619179 (2009).
14. Yankeelov, T. E., Rooney, W. D., Li, X. & Springer, C. S. Jr. Variation of the relaxographic "shutter-speed" for transcytolemmal water exchange affects the CR bolus-tracking curve shape. *Magnetic resonance in medicine* **50**, 1151–1169, https://doi.org/10.1002/mrm.10624 (2003).
15. Leach, M. *et al.* Imaging vascular function for early stage clinical trials using dynamic contrast-enhanced magnetic resonance imaging. *European radiology* **22**, 1451–1464 (2012).
16. Schabel, M. C. & Parker, D. L. Uncertainty and bias in contrast concentration measurements using spoiled gradient echo pulse sequences. *Physics in medicine and biology* **53**, 2345–2373, https://doi.org/10.1088/0031-9155/53/9/010 (2008).

17. Yang, C. *et al*. Reproducibility assessment of a multiple reference tissue method for quantitative dynamic contrast enhanced-MRI analysis. *Magnetic resonance in medicine* **61**, 851–859, https://doi.org/10.1002/mrm.21912 (2009).
18. Heisen, M. *et al*. The influence of temporal resolution in determining pharmacokinetic parameters from DCE-MRI data. *Magnetic resonance in medicine* **63**, 811–816, https://doi.org/10.1002/mrm.22171 (2010).
19. Di Giovanni, P. *et al*. The accuracy of pharmacokinetic parameter measurement in DCE-MRI of the breast at 3 T. *Physics in medicine and biology* **55**, 121–132, https://doi.org/10.1088/0031-9155/55/1/008 (2010).
20. Sourbron, S. P. & Buckley, D. L. On the scope and interpretation of the Tofts models for DCE-MRI. *Magnetic resonance in medicine* **66**, 735–745, https://doi.org/10.1002/mrm.22861 (2011).
21. Sourbron, S. P. & Buckley, D. L. Tracer kinetic modelling in MRI: estimating perfusion and capillary permeability. *Physics in medicine and biology* **57**, R1–33, https://doi.org/10.1088/0031-9155/57/2/R1 (2012).
22. Othman, A. E. *et al*. Comparison of different population-averaged arterial-input-functions in dynamic contrast-enhanced MRI of the prostate: Effects on pharmacokinetic parameters and their diagnostic performance. *Magnetic resonance imaging* **34**, 496–501, https://doi.org/10.1016/j.mri.2015.12.009 (2016).
23. Tofts, P. S. & Kermode, A. G. Measurement of the blood-brain barrier permeability and leakage space using dynamic MR imaging. 1. Fundamental concepts. *Magnetic resonance in medicine* **17**, 357–367 (1991).
24. Heye, T. *et al*. Reproducibility of dynamic contrast-enhanced MR imaging. Part I. Perfusion characteristics in the female pelvis by using multiple computer-aided diagnosis perfusion analysis solutions. *Radiology* **266**, 801–811, https://doi.org/10.1148/radiol.12120278 (2013).
25. Huang, W. *et al*. Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge. *Translational oncology* **7**, 153–166 (2014).
26. Beuzit, L. *et al*. Dynamic contrast-enhanced MRI: Study of inter-software accuracy and reproducibility using simulated and clinical data. *Journal of magnetic resonance imaging* **43**, 1288–1300, https://doi.org/10.1002/jmri.25101 (2016).
27. Cron, G. O. *et al*. Bias and precision of three different DCE-MRI analysis software packages: a comparison using simulated data. *International Society for Magnetic Resonance in Medicine*. (Proc 22nd Annual Meeting ISMRM, Milan (abstract 4592)) (2014).
28. Tofts, P. S. Modeling tracer kinetics in dynamic Gd-DTPA MR imaging. *Journal of magnetic resonance imaging* **7**, 91–101 (1997).
29. Tofts, P. S. *et al*. Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols. *Journal of magnetic resonance imaging* **10**, 223–232 (1999).
30. *Quantitative Imaging Biomarkers Alliance*, https://www.rsna.org/QIBA/.
31. Carletta, J. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics* **22**, 249–254 (1996).
32. Krippendorff, K. Reliability in content analysis. *Human communication research* **30**, 411–433 (2004).
33. Neuendorf, K. A. *The content analysis guidebook*. (Sage, 2002).
34. Kim, S. *et al*. Prediction of response to chemoradiation therapy in squamous cell carcinomas of the head and neck using dynamic contrast-enhanced MR imaging. *American journal of neuroradiology* **31**, 262–268, https://doi.org/10.3174/ajnr.A1817 (2010).
35. Van Cann, E. M. *et al*. Quantitative dynamic contrast-enhanced MRI for the assessment of mandibular invasion by squamous cell carcinoma. *Oral oncology* **44**, 1147–1154, https://doi.org/10.1016/j.oraloncology.2008.02.009 (2008).
36. Lee, F. K., King, A. D., Ma, B. B. & Yeung, D. K. Dynamic contrast enhancement magnetic resonance imaging (DCE-MRI) for differential diagnosis in head and neck cancers. *European journal of radiology* **81**, 784–788, https://doi.org/10.1016/j.ejrad.2011.01.089 (2012).
37. Bisdas, S. *et al*. An exploratory pilot study into the association between microcirculatory parameters derived by MRI-based pharmacokinetic analysis and glucose utilization estimated by PET-CT imaging in head and neck cancer. *European radiology* **20**, 2358–2366, https://doi.org/10.1007/s00330-010-1803-x (2010).
38. Tofts, P. S., Berkowitz, B. & Schnall, M. D. Quantitative analysis of dynamic Gd-DTPA enhancement in breast tumors using a permeability model. *Magnetic resonance in medicine* **33**, 564–568 (1995).
39. Ashton, E. Quantitative MR in multi-center clinical trials. *Journal of magnetic resonance imaging* **31**, 279–288, https://doi.org/10.1002/jmri.22022 (2010).
40. Schabel, M. C. & Morrell, G. R. Uncertainty in T(1) mapping using the variable flip angle method with two flip angles. *Physics in medicine and biology* **54**, N1–8, https://doi.org/10.1088/0031-9155/54/1/N01 (2009).
41. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 7900–7905, https://doi.org/10.1073/pnas.1602413113 (2016).
42. Galbraith, S. M. *et al*. Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumours: comparison of quantitative and semi-quantitative analysis. *NMR in biomedicine* **15**, 132–142 (2002).
43. Butterworth, E., Jardine, B. E., Raymond, G. M., Neal, M. L. & Bassingthwaighte, J. B. JSim, an open-source modeling system for data analysis. *F1000Research* **2**, 288, https://doi.org/10.12688/f1000research.2-288.v1 (2013).
44. Barboriak, D. P. QIBA_v6_Tofts_RevB, https://sites.duke.edu/dblab/qibacontent/.
45. Barboriak, D. P. QIBA_v9_Tofts, https://sites.duke.edu/dblab/qibacontent/.
46. Gelman, A. & Hill, J. *Data analysis using regression and multilevel/hierarchical models*. 45–46 (Cambridge University Press, 2006).
47. Korporaal, J. G. *et al*. Phase-based arterial input function measurements in the femoral arteries for quantification of dynamic contrast-enhanced (DCE) MRI and comparison with DCE-CT. *Magnetic resonance in medicine* **66**, 1267–1274, https://doi.org/10.1002/mrm.22905 (2011).
48. *NordicNeuroLab*, http://www.nordicneurolab.com/.
49. Li, X. *et al*. Dynamic NMR effects in breast cancer dynamic-contrast-enhanced MRI. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 17937–17942, https://doi.org/10.1073/pnas.0804224105 (2008).
50. Tudorica, A. *et al*. Early Prediction and Evaluation of Breast Cancer Response to Neoadjuvant Chemotherapy Using Quantitative DCE-MRI. *Translational oncology* **9**, 8–17, https://doi.org/10.1016/j.tranon.2015.11.016 (2016).
51. Coolens, C. *et al*. Automated voxel-based analysis of volumetric dynamic contrast-enhanced CT data improves measurement of serial changes in tumor vascular biomarkers. *International journal of radiation oncology, biology, physics* **91**, 48–57, https://doi.org/10.1016/j.ijrobp.2014.09.028 (2015).
52. Coolens, C., Driscoll, B., Moseley, J., Brock, K. K. & Dawson, L. A. Feasibility of 4D perfusion CT imaging for the assessment of liver treatment response following SBRT and sorafenib. *Advances in Radiation Oncology* **1**, 194–203 (2016).
53. Hormuth, D. A. 2nd, Skinner, J. T., Does, M. D. & Yankeelov, T. E. A comparison of individual and population-derived vascular input functions for quantitative DCE-MRI in rats. *Magnetic resonance imaging* **32**, 397–401, https://doi.org/10.1016/j.mri.2013.12.019 (2014).
54. Barnes, S. L., Whisenant, J. G., Loveless, M. E. & Yankeelov, T. E. Practical dynamic contrast enhanced MRI in small animal models of cancer: data acquisition, data analysis, and interpretation. *Pharmaceutics* **4**, 442–478, https://doi.org/10.3390/pharmaceutics4030442 (2012).


## Acknowledgements

Multiple funders/agencies contributed to contributing personnel salary or project support during the study execution and manuscript preparation interval. Drs. Lai, Mohamed and Fuller receive funding support from

the National Institutes of Health (NIH)/National Institute for Dental and Craniofacial Research (NIDCR) (R01DE025248). Dr. Fuller received/(s) grant and/or salary support from the NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Career Development Award (P50CA097007-10); the NCI Paul Calabresi Clinical Oncology Program Award (K12 CA088084-06); a General Electric Healthcare/MD Anderson Center for Advanced Biomedical Imaging In-Kind Award; an Elekta AB/MD Anderson Department of Radiation Oncology Seed Grant; the Center for Radiation Oncology Research (CROR) at MD Anderson Cancer Center Seed Grant; the MD Anderson Institutional Research Grant (IRG) Program; and the NIH/NCI Cancer Center Support (Core) Grant CA016672 to The University of Texas MD Anderson Cancer Center (P30 CA016672). Rachel Ger is supported by the Rosalie B. Hite Graduate Fellowship in Cancer Research awarded by The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences. Dr. Huang is funded by the NIH (U01 CA154602). Drs. Hormuth and Yankeelov are supported by the NIH (R01CA138599, U01CA174706) and Cancer Prevention Research Institute of Texas (CPRIT RR160005).

## Author Contributions

R.B.G.- Drafted manuscript, performed supervised image analysis, statistical analysis, contributed important intellectual modification of initial study concept. A.S.R.M.-Study coordinator; assisted with study conception; oversaw all image processing, image-registration, clinical data collection, prospective patient enrollment and administrative oversight; responsible for trainee (R.B.G., M.J.A., H.E.) and clinical/research staff oversight (Y.D). M.J.A.-Responsible for custom software for image-registration of clinical datasets, drafted portions of clinical MRI acquisition protocol under supervision, assisted with initial conception. Y.D., J.W., R.J.S., V.C.S. - Responsible for prospective clinical dataset image acquisition and quality assurance, pre-analytic processing and image transmission between institutions. S.Z.- Responsible for overseeing statistical tests, creating stratified permutation test; supervision of trainee (R.B.G). R.M.H., H.L.- Responsible for mentored oversight of trainee (R.B.G.); manuscript assistance. L.E.C.- Responsible for statistical analysis and interpretation, manuscript drafting, manuscript review and approval; supervision of trainee (R.B.G). K.L., X.J.F., A.L.B., B.D., H.E., D.A.H., P.J.v.H., R.H., K.B.M., J.A.B., W.H., T.Y., U.A.v.d.H.- Direct contribution of data, expertise, and image-analysis of study data; custom software construction and implementation; data dissemination and oversight. D.P.B.- Responsible for initial conceptualization of utilized digital reference object, specific expertise contribution, as well as DRO data provision. C.C., C.C.- Programmatic development and trial coordination assistance and project support; manuscript and study design assistance. S.Y.L.- Programmatic development and trial coordination assistance and project support; manuscript and study design assistance; oversight and direct funding of clinical/research coordination personnel (A.S.R.M). S.J.F.- Parent clinical trial primary investigator; direct provision of clinical research infrastructure for image acquisition, patient data collection, and approval for use. Responsible for mentored oversight of clinical/research personnel and direct clinical trial development mentorship (C.D.F). J.D.H.-Direct clinical imaging programmatic oversight, resource provision, and mentored iterative protocol development; direct imaging informatics mentorship (C.D.F). J.K.C.-Study co-director; co-conceived of study with C.D.F.; performed direct software analysis, assisted with multi-site quality assurance, study design oversight, as well as implementation of digital reference object analytics. C.D.F.- Study initiator and co-director; co-conceived of study with J.K.C., oversight of all portions of study, including clinical data acquisition, patient data collection, pre-and post-processing, study coordination, statistical analysis and interpretation, data dissemination, manuscript drafting, manuscript review and approval; supervision of trainee (R.B.G., M.J.A., H.E.) and clinical/research staff (A.S.R.M., Y.D).

## Additional Information

**Supplementary information** accompanies this paper at doi:10.1038/s41598-017-11554-w

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Consortia
## Joint Head and Neck Radiotherapy-MRI Development Cooperative

**Rachel B. Ger[1,2], Abdallah S. R. Mohamed[3,4], Musaddiq J. Awan[5,6], Yao Ding[7], Kimberly Li[8,9], Xenia J. Fave[1,2], Andrew L. Beers[10], Brandon Driscoll[11], Hesham Elhalawani[3], David A. Hormuth II[12], Petra J. van Houdt[13], Renjie He[14], Shouhao Zhou[15], Kelsey B. Mathieu[7], Heng Li[1,2], Catherine Coolens[11,16,17], Caroline Chung[3,11], James A. Bankson[2,7], Wei Huang[8], Jihong Wang[1,2], Vlad C. Sandulache[18], Stephen Y. Lai[19,20], Rebecca M. Howell[1,2], R. Jason Stafford[2,7], Thomas E. Yankeelov[12], Uulke A. van der Heide[13], Steven J. Frank[3], Daniel P. Barboriak[21], John D. Hazle[2,7], Laurence E. Court[1,2,7], Jayashree Kalpathy-Cramer[10] & Clifton D. Fuller[2,3]**

[1]Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [2]The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, Texas, USA. [3]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [4]Department of Clinical Oncology and Nuclear Medicine, Faculty of Medicine, University of Alexandria, Alexandria, Egypt. [5]Case Western Reserve University, Cleveland, OH, USA. [6]University Hospitals, Cleveland, OH, USA. [7]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [8]Advanced Imaging Research Center, Knight Cancer Institute, Oregon Health & Science University, Portland, Oregon, USA. [9]The International School of Beaverton, Beaverton, Oregon, USA. [10]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital/Division of Health Sciences & Technology, Massachusetts Institute of Technology, Charlestown, Massachusetts, USA. [11]Techna Institute, University Health Network, Toronto, Ontario, Canada. [12]Institute for Computational Engineering and Sciences, The University of Texas, Austin, Texas, USA. [13]Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [14]United Imaging Healthcare America, Houston, Texas, USA. [15]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [16]Radiation Medicine Program, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [17]Department of Radiation Oncology, University of Toronto, Toronto, Ontario, Canada. [18]Department of Otolaryngology Head and Neck Surgery, Baylor College of Medicine, Houston, Texas, USA. [19]Department of Head and Neck Surgery, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [20]Department of Molecular and Cellular Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [21]Department of Radiology, Duke University Medical Center, Durham, North Carolina, USA.