Bias and precision of three different DCE-MRI analysis software packages: a comparison using simulated data

Greg O Cron[1,2], Steven Sourbron[3], Daniel P Barboriak[4], Rhys Abdeen[2], Matthew Hogan[2,5], and Thanh B Nguyen[1,2]

[1]*Medical Imaging, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada,* [2]*University of Ottawa, Ottawa, Ontario, Canada,* [3]*Institute of Genetics, Health, and Therapeutics, University of Leeds, Leeds, United Kingdom,* [4]*Radiology, Duke University Medical Center, Durham, North Carolina, United States,* [5]*Neuroscience, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada*

Purpose: Software packages (SPs) for analysis of DCE-MRI data abound. However, it is unclear how consistent they are with each other (even for supposedly identical models), which impedes standardization. Heye et al investigated this issue in Radiology 266:801, testing inter-software reproducibility on uterine fibroid DCE-MRI data. However, in terms of comparing SPs, the Heye study was confounded by having multiple observers and variable regions of interest. Moreover, for clinical data the underlying ("ground truth") tracer kinetic parameters are unknown, precluding any investigation of bias (difference from ground truth). The purpose of this work was to compare three DCE-MRI SPs, using identical simulated data, to see if the SPs have different bias and precision.

Methods: The three SPs performed pixel-by-pixel image analysis of DCE-MRI data, using the Extended Tofts model (ETM). All SPs used the Murase linear fit method to derive parameters from ETM (MRM 51:858). One of the SPs was freeware developed by a leading academic DCE-MRI researcher, whereas the other two were commercial. The same noiseless, simulated image data were input into all three SPs. These data had been generated from ETM (JMRI 27:1388) with different combinations of the volume transfer coefficient (Ktrans = 0 to 0.2 min^{-1}), plasma volume fraction (vp = 0.001 to 0.1), and distribution volume fraction (ve = 0.1 to 0.5). The data were organized as a time series of integer-valued DICOM images, with image intensities equal to the gadolinium concentration (in units of μM) rounded to the nearest integer (perfect T1 mapping and signal conversion assumed). The total acquisition time (Tacq) was 5.5 minutes and Δt = 3.5 s, the latter chosen to match our clinical protocol (AJNR 33:1539). The matrix size of the images was 60 x 200, with different pixels representing different combinations of Ktrans, vp, and ve. For each SP, a 3 x 3-pixel region of interest was placed in a purely vascular region, providing the arterial input function. To observe the effect of signal noise on the SPs, we repeated the study, adding noise to the signal data (before conversion to μM of Gad) at a typical level which we observe in our clinic for DCE-MRI of gliomas: 4% of baseline signal (=2% of peak change in signal). We repeated again with "high" noise (2.5 x greater). Any fit which returned any unphysical value (UV, i.e. negative or fraction > 1) was excluded from further analysis. These SPs did not return any NaN values.

Results: For all SPs, % UVs increased with noise and were smallest for Ktrans (except Ktrans=0.01 min^{-1} at high noise). Moreover, precision was usually worse for ve. There also appeared to be a trend toward lower returned values at higher noise, which was particularly evident with SP B. SPs A and C gave very similar results, except at zero noise where SP A had significant numbers of UVs.

	zero noise			typical noise			high noise		
	A	B	C	A	B	C	A	B	C
Ktrans=0.01	1.16 ± 0.07	1.16 ± 0.07	1.16 ± 0.07	1.03 ± 0.17	1.00 ± 0.17	1.04 ± 0.17	1.04 ± 0.32	0.91 ± 0.36	1.04 ± 0.34
Ktrans=0.02	1.09 ± 0.04	1.09 ± 0.04	1.09 ± 0.04	1.02 ± 0.10	0.99 ± 0.10	1.02 ± 0.10	0.94 ± 0.23	0.84 ± 0.22	0.95 ± 0.23
Ktrans=0.05	1.04 ± 0.02	1.04 ± 0.02	1.04 ± 0.02	1.03 ± 0.06	0.99 ± 0.06	1.03 ± 0.06	0.97 ± 0.14	0.86 ± 0.13	0.97 ± 0.14
Ktrans=0.10	1.03 ± 0.01	1.03 ± 0.01	1.03 ± 0.01	1.03 ± 0.04	0.99 ± 0.04	1.03 ± 0.04	1.00 ± 0.10	0.88 ± 0.09	1.01 ± 0.10
Ktrans=0.20	1.01 ± 0.03	1.02 ± 0.02	1.02 ± 0.02	1.03 ± 0.04	0.98 ± 0.03	1.03 ± 0.04	1.02 ± 0.08	0.88 ± 0.08	1.03 ± 0.09
vp=0.005	1.14 ± 0.08	1.14 ± 0.08	1.14 ± 0.08	1.06 ± 0.25	0.98 ± 0.23	1.05 ± 0.25	0.90 ± 0.58	0.68 ± 0.49	0.87 ± 0.56
vp=0.010	1.08 ± 0.04	1.07 ± 0.04	1.07 ± 0.04	1.04 ± 0.13	0.97 ± 0.12	1.04 ± 0.13	0.96 ± 0.30	0.78 ± 0.26	0.95 ± 0.30
vp=0.020	1.04 ± 0.01	1.04 ± 0.02	1.04 ± 0.02	1.04 ± 0.07	0.97 ± 0.07	1.04 ± 0.07	0.98 ± 0.17	0.81 ± 0.15	0.98 ± 0.17
vp=0.050	1.02 ± 0.01	1.02 ± 0.01	1.02 ± 0.01	1.04 ± 0.04	0.98 ± 0.04	1.04 ± 0.04	1.01 ± 0.10	0.84 ± 0.08	1.01 ± 0.10
vp=0.100	1.02 ± 0.00	1.02 ± 0.00	1.02 ± 0.00	1.04 ± 0.03	0.98 ± 0.03	1.04 ± 0.04	1.02 ± 0.08	0.85 ± 0.07	1.02 ± 0.08
ve=0.1	1.06 ± 0.12	1.05 ± 0.59	1.05 ± 0.58	1.03 ± 0.18	1.00 ± 0.16	1.04 ± 0.17	0.98 ± 0.22	0.89 ± 0.20	0.98 ± 0.22
ve=0.2	1.04 ± 0.13	1.04 ± 0.55	1.04 ± 0.55	1.03 ± 0.22	0.99 ± 0.25	1.03 ± 0.22	0.99 ± 0.23	0.87 ± 0.21	0.99 ± 0.21
ve=0.5	1.02 ± 0.18	0.99 ± 0.51	1.00 ± 0.51	1.04 ± 0.29	0.98 ± 0.38	1.03 ± 0.25	1.02 ± 0.30	0.87 ± 0.43	1.00 ± 0.24

Table 1: Median ± precision of values returned by SPs A, B, and C, normalized to expected values shown in left column (perfection = 1.00±0.00), excluding fits containing UVs. Precision was defined as ½ (10^{th} – 90^{th} percentile). % of fits with UVs: 0-5; 6-10; 11-20; 21-50; 51-100.

Discussion: We observed differences between the SPs, especially in terms of how they react to noise. This was somewhat surprising, considering that all three SPs used supposedly identical algorithms (ETM + Murase linear fit). The synthetic data had been designed to cover physiological situations, however certain parameter spaces edged into non-physiological territory, where ETM fitting algorithms might be inherently unstable. It is therefore possible that the differences between the SPs may be due to their varied reactions to these non-physiological sections. The generally poor precision of the SPs for measuring ve may reflect the fact that if Tacq is shorter than the interstitial MTT=ve/Ktrans, then ve is inherently unmeasurable – no matter how good the software is. These results support the notion that the bias and precision of a DCE-MRI analysis may be affected by the particular SP used. It also must be kept in mind that there are other factors to consider beyond strict bias and precision (e.g. practicality in clinical routine, calculation times, cost, level of expertise required, level of manual intervention, and support).

Conclusion: Ideally, SPs should not be mixed, i.e. a single SP should be used for a given DCE-MRI study. If inter-SP comparison cannot be avoided (e.g. retrospective inter-institutional comparison), then investigators should rigorously characterize any systematic SP differences.