Comparison of algorithms for linear-combination modelling of short-echo-time magnetic resonance spectra

Helge J. Zöllner[1,2,*], Michal Považan[1,2], Steve C. N. Hui[1,2], Sofie Tapper[1,2], Richard A. E. Edden[1,2], Georg Oeltzschner[1,2]

[1] Russell H. Morgan Department of Radiology and Radiological Science, The Johns Hopkins University School of Medicine, Baltimore, MD, United States

[2] F. M. Kirby Research Center for Functional Brain Imaging, Kennedy Krieger Institute, Baltimore, MD, United States

***Corresponding author:**

Georg Oeltzschner, Ph.D.

Division of Neuroradiology, Park 367G

The Johns Hopkins University School of Medicine

600 N Wolfe St

Baltimore, MD 21287

goeltzs1@jhmi.edu

Running title: Comparison of MRS linear-combination algorithms

Word count: 5116

Keywords

- Magnetic resonance spectroscopy
- Linear-combination modelling
- Short-TE PRESS

Highlights

- Means and CVs for tNAA and tCho are highly consistent across vendors and algorithms.
- Means and CVs for mI and Glx are less consistent across vendors and algorithms.
- Agreement between metabolite estimates from different algorithms is moderate at best.
- Baseline estimation contributes significantly to measurement variance.

Abstract

Short-TE proton magnetic resonance spectroscopy is commonly used to study metabolism in the human brain. Spectral modelling is a crucial analysis step, yielding quantitative estimates of metabolite levels. Commonly used quantification methods model the data as a linear combination of metabolite basis spectra, maximizing the use of prior knowledge to constrain the model solution. Various linear-combination modelling (LCM) algorithms have been integrated into widely used commercial and open-source analysis programs.

This large-scale, in-vivo, multi-vendor study compares the levels of four major metabolite complexes estimated by three LCM algorithms (Osprey, LCModel, and Tarquin). 277 short-TE spectra from a recent multi-site study were pre-processed with the Osprey software. The resulting spectra were modelled with Osprey, Tarquin and LCModel, using the same vendor-specific basis sets. Levels of total N-acetylaspartate (tNAA), total choline (tCho), myo-inositol (mI), and glutamate+glutamine (Glx) were quantified with respect to total creatine (tCr).

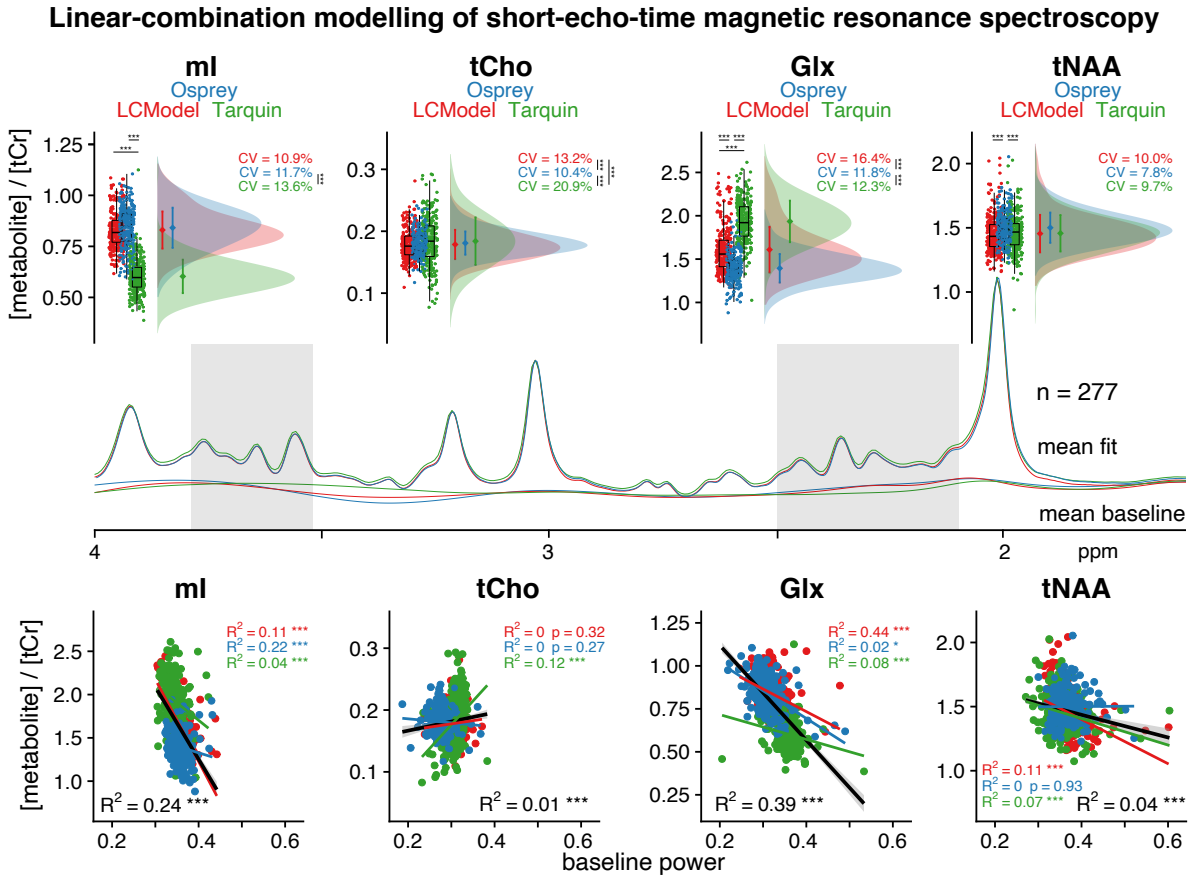
Mean spectra and models showed high agreement between all vendors and LCM algorithms. In contrast, the algorithms differed notably in their baseline estimates and mI models. Group means and CVs of the metabolite estimates agreed well for tNAA and tCho across vendors and algorithms. For mI and Glx, group means metabolite estimates and CVs agreed less well between algorithms, with mI systematically estimated lower by Tarquin. Across all datasets, the metabolite-mean $\overline{CV}$ was 10.4% for Osprey, 12.6% for LCModel, and 14.0% for Tarquin. The grand mean correlation coefficient for all pairs of LCM algorithms across all datasets and metabolites was $\overline{R^2} = 0.39$, indicating generally moderate agreement of individual metabolite estimates between algorithms. Stronger correlations were found for tNAA and tCho than for Glx and mI.

Correlations between pairs of algorithms were comparably moderate (Tarquin-vs-LCModel: $\overline{R^2}$ = 0.43; Osprey-vs-LCModel: $\overline{R^2}$ = 0.38; Osprey-vs-Tarquin: $\overline{R^2}$ = 0.37). There was a significant association between local baseline power and metabolite estimates (grand mean $\overline{R^2}$ = 0.10; up to 0.62 for LCModel analysis of Glx in Siemens datasets). Metabolite estimates with stronger associations to the local baseline power (mI and Glx) showed higher variance, suggesting that baseline estimation has a stronger influence on those metabolites than on tNAA and tCho.

While estimates of major metabolite complexes broadly agree between linear-combination modelling algorithms at group level, correlations between algorithms are only weak to moderate, despite standardized pre-processing, a large sample of young, healthy and cooperative subjects, and high spectral quality. These findings raise concerns about the comparability of MRS studies, which typically use one algorithm and much smaller sample sizes.

The study is the first benchmark comparison of different LCM algorithms using large datasets. Finally, it provides a standardized framework to interrogate factors with critical impact on LCM outcomes.

Graphical Abstract



Introduction

Proton MR spectroscopy (^{1}H-MRS) allows in-vivo clinical and basic research studies of the metabolism (Öz et al., 2014; Wilson et al., 2019). Single-voxel MR spectra from the human brain are frequently acquired using point-resolved spectroscopy (PRESS) localization (Bottomley, 1985) , and can be modelled to estimate metabolite levels. Accurate modelling is hampered by poor spectral resolution at clinical field strengths, and for short-echo-time spectra metabolite signals overlap with a broad background consisting of fast-decaying macromolecule and lipid signals. Linear-combination modelling (LCM) of the spectra maximizes the use of prior knowledge to constrain the model solution, and was recommended by recent consensus (Near et al., 2020). LCM algorithms model spectra as a linear combination of (metabolite and macromolecular (MM)) basis functions, and typically also include terms to account for smooth baseline fluctuations.

Several LCM algorithms are available to quantify MR spectra (**Table 1** describes some of the most widely used), either as open-source (Osprey (Oeltzschner et al., 2020), Tarquin (Wilson et al., 2011), Vespa (Soher et al., 2011)) or compiled 'black-box' implementations (LCModel (Provencher, 1993), QUEST (Graveron-Demilly, 2014), AQSES (Poullet et al., 2007), INSPECTOR (Juchem, 2018)). Most LCM algorithms are freely available for academic purposes (Osprey, Tarquin, QUEST and AQSES in jMRUI (Naressi et al., 2001; Stefan et al., 2009), Vespa, INSPECTOR). Algorithms operate in either the time (Tarquin, QUEST, AQSES) or frequency domain (LCModel, Osprey, INSPECTOR), and use different approaches to address the macromolecular background. The smooth baseline can either be constructed from cubic B-spline

89 functions (LCModel, Osprey, QUEST, AQSES), by truncating the initial FID and filtering the

90 residual (Tarquin), from a set of wavelets (Vespa), or as a first-order polynomial (INSPECTOR).

Name	**Modelling Domain, Baseline approach**	**Cost**	**Code Availability**	**Published**	**Cita-tions***
Osprey	FD, spline baseline	free	open	2020	0
LCModel	FD, spline baseline	$13,300	closed	1993	3384
Tarquin	TD, smooth baseline	free	open	2011	243
QUEST (jMRUI)	TD, spline baseline	free	closed	2004	305
AQSES (jMRUI)	TD, spline baseline	free	closed	2007	136
Vespa	FD, wavelet baseline	free	open	2006	66
INSPECTOR	FD, 1st-order polynomial	free	closed	2018	0

91 Table 1 – Overview of linear-combination modelling algorithms. The domain (either time TD or
92 frequency FD) of modelling and the baseline model approach is specified. *Citations re-
93 ported from Google Scholar on May 7 2020.
94

95 Surprisingly few studies have compared the performance of different LCM algorithms. Cross-

96 validation of quantitative results has almost exclusively been performed in the context of bench-

97 marking new algorithms against existing solutions. In-vivo comparisons are often limited to

98 small sample sizes, whether analyzing spectra from animal models (Osorio-Garcia et al., 2011;

99 Shen et al., 2008; Wilson et al., 2011) or human subjects (Osorio-Garcia et al., 2011; Poullet et

100 al., 2007; Wilson et al., 2011). To our knowledge, only (Kossowski et al., 2017) in rat brain and

101 (Mosconi et al., 2014) in human body have compared the LCM performance of different algo-

102 rithms. Scott et al., 2016 compares LCM algorithms to non-LCM quantification, such as

103 AMARES (Vanhamme et al., 1997), observing consistent differences. Most studies report good

104 agreement between results from different algorithms, inferring this from group-mean compari-

105 sons, or observing that differences between clinical groups are consistent regardless of the algo-

106 rithm applied (Kossowski et al., 2017; Scott et al., 2016). Correlations of estimates from

different algorithms are rarely reported; (Kossowski et al., 2017) shows a high correlation between LCModel and Tarquin results from ultra-high-field MRS data.

Another major group of studies focuses on mechanistic investigations of the LCM process itself. These report that metabolite estimates are strongly influenced by data quality, both signal-to-noise ratio (SNR) and linewidth (Bartha, 2005; Wijtenburg and Knight-Scott, 2009; Zhang and Shen, 2020). Critical modelling decisions also impact outcomes, e.g. measured or simulated macromolecular spectra (Marjańska and Terpstra, 2019; Schaller et al., 2014), measured or simulated basis sets (Cudalbu et al., 2008), and baseline modelling approach (Gottschalk et al., 2008), including the baseline smoothness (Marjańska and Terpstra, 2019; Wenger et al., 2019). In summary, despite the fact that LCM has been used to analyze thousands of studies (Table 1), the process itself has been studied to surprisingly small extent. A comprehensive assessment of the agreement between the algorithms is lacking, and the relationship between the choice of model parameters and quantitative outcomes is poorly understood.

To address this gap, we conducted a large-scale comparison of short-TE in-vivo MRS data using three LCM algorithms with standardized pre-processing. We recently developed 'Osprey' (https://github.com/schorschinho/osprey, 2020), an open-source MATLAB (The MathWorks Inc., Natick, MA) toolbox for pre-processing and quantification of MR spectra. Osprey includes functionality for the export of basis sets for further use in third-party software and of pre-processed data prior to LCM, making it a useful framework to interrogate the performance of the LCM algorithms in LCModel, Tarquin and Osprey. We compared group-mean quantification results of four major metabolite complexes from each LCM algorithm, performed between-algorithm correlation analyses, and investigated local baseline power and creatine modelling as

potential sources of differences between the algorithms, which may improve our mechanistic understanding of the LCM process itself.

Methods

Study participants & data acquisition

277 single-voxel short-TE PRESS datasets from healthy volunteers acquired in a recent multi-site-study (Považan et al., 2020) were included in this analysis. A subset of these data is available at (https://www.nitrc.org/projects/biggaba/, 2018). Data were acquired at 25 sites (with up to 12 subjects per site) on scanners from three different vendors (GE, Philips, and Siemens) with the following parameters: TR/TE = 2000/35 ms; 64 averages; 2, 4 or 5 kHz spectral bandwidth: 2048-4096 data points; acquisition time = 2.13 min; $3 \times 3 \times 3$ cm^3 voxel in the medial parietal lobe). Other acquisition parameters (including water suppression, phase cycling, outer volume suppression, and B_0 homogeneity adjustment) were chosen according to local preference. Reference spectra were acquired with similar parameters, but without water suppression and 8-16 averages (for more details, please refer to (Považan et al., 2020)). Raw data were exported from the scanner in vendor-native formats (GE P-files, Philips .sdat, and Siemens TWIX). Written informed consent was obtained from each participant and the study was approved by the local institutional review boards.

Data pre-processing

MRS data were pre-processed in Osprey (Oeltzschner et al., 2020), an open-source MATLAB toolbox, following recent peer-reviewed pre-processing recommendations (Wilson et al., 2019). The three steps will be briefly described in the following and are summarized in **Figure 1A**. First, using the *OspreyLoad* module, the vendor-native raw data were loaded, including the metabolite (water-suppressed) data, as well as the unsuppressed water reference data. Second, the *OspreyProcess* module transformed the raw data into pre-processed, averaged spectra. Receiver-

coil combination (Hall et al., 2014) and eddy-current correction (Klose, 1990) of the metabolite data were performed using the water reference data. Individual transients in Siemens and GE data were frequency-and-phase aligned using robust spectral registration (Mikkelsen et al., 2019), while Philips data had been averaged on the scanner. After averaging the individual transients, the residual water signal in the metabolite spectra was removed with a Hankel singular value decomposition filter (Barkhuijsen et al., 1987). For Siemens spectra, an additional pre-phasing step was introduced by modelling the signals from creatine and choline-containing compounds at 3.02 and 3.20 ppm with a double Lorentzian model and applying the phase of the model to the spectra. Third, the pre-processed spectra were exported in .RAW format for LCModel and Tarquin.

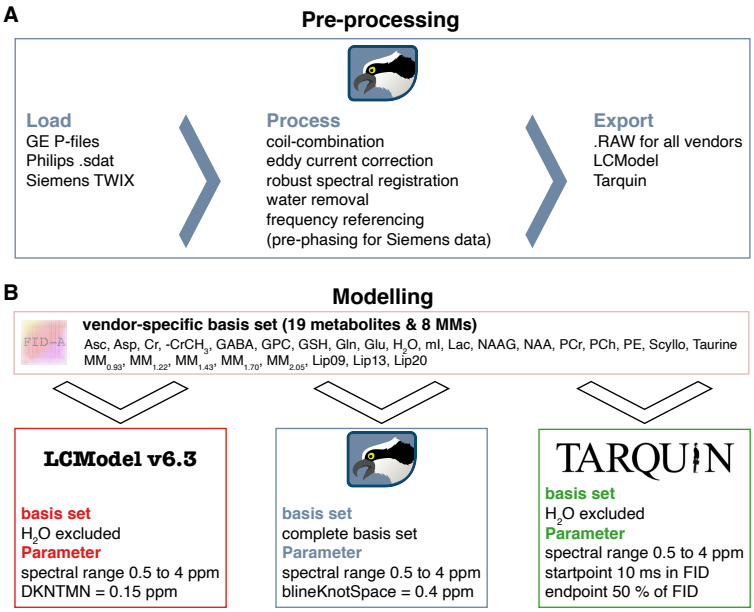


Figure 1– Overview of the MRS analysis pipeline. (A) Pre-processing pipeline implemented in Osprey including 'OspreyLoad' to load the vendor-native spectra, 'OspreyProcess' to process the raw data and to export the averaged spectra. (B) Modelling of the averaged spectra with details of the basis set and parameters of each LCM (LCModel, Osprey, and Tarquin).

Data modelling

Density-matrix simulations implemented in the MATLAB toolbox FID-A (Simpson et al., 2017) were used to generate vendor-specific basis sets including 19 spin systems: ascorbate, aspartate, creatine (Cr), negative creatine methylene (-$CrCH_2$), γ-aminobutyric acid (GABA), glycerophosphocholine (GPC), glutathione, glutamine (Gln), glutamate (Glu), water (H_2O), myo-inositol (mI), lactate, N-acetylaspartate (NAA), N-acetylaspartylglutamate (NAAG), phosphocholine (PCh), phosphocreatine (PCr), phosphoethanolamine, scyllo-inositol, and taurine. 8 additional Gaussian basis functions were included in the basis set to simulate broad macromolecules and lipid resonances using Gaussian signals following the consensus nomenclature (Cudalbu et al., 2020) (simulated as defined in section 11.7 of the LCModel manual): $MM_{0.94}$, $MM_{1.22}$, $MM_{1.43}$, $MM_{1.70}$, $MM_{2.05}$, Lip09, Lip13, Lip20. Finally, to standardize the basis set for each algorithm, basis sets were stored as .mat files for use in Osprey and as .BASIS-files for use in LCModel and Tarquin. In the following, each LCM algorithm investigated in this study will be described briefly (for details, please refer to the original publications).

LCModel v6.3

The LCModel (6.3-0D) algorithm (Provencher, 1993) models data in the frequency-domain. First, time-domain data and basis functions are zero-filled by a factor of two. Second, frequency-domain spectra are frequency-referenced by cross-correlating them with a set of delta functions representing the major singlet landmarks of NAA (2.01 ppm), Cr (3.02 ppm), and Cho (3.20 ppm). Third, starting values for phase and linebroadening parameters are estimated by modelling the data with a reduced basis set containing NAA, Cr, PCh, Glu, and mI, with a smooth baseline. Fourth, the final modelling of the data is performed with the full basis set, regularized lineshape model and baseline, with starting values for phase, linebroadening, and lineshape parameters

derived from the previous step. Model parameters are determined with a Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963) non-linear least-squares optimization implementation that allows hard constraints to be imposed on the parameters. Metabolite amplitudes are constrained to be non-negative, and determined using a non-negative linear least-squares fit at each iteration of the non-linear optimization. Weak soft constraints on macromolecule and lipid amplitudes - and select pairs of metabolites (e.g. NAA and NAAG) - are defined as in Osprey and Tarquin. In contrast to Osprey, LCModel constrains the model with three additional regularization terms, to penalize a lack of smoothness in the baseline and lineshape models, and deviations of the Lorentzian linebroadening and frequency shift parameters from their expectation values.

Osprey

The Osprey (1.0.0) algorithm (Oeltzschner et al., 2020) adopts several key features of the LCModel and Tarquin algorithms. Osprey models the real part of the frequency-domain spectrum. First, time-domain data and basis functions are zero-filled by a factor of two and zero-order phase correction, global Gaussian linebroadening, individual Lorentzian linebroadening, and individual frequency shifts are applied to each basis function before Fourier transformation. The frequency-domain basis functions are then convolved with an arbitrary, unregularized lineshape model to account for deviations from a Voigtian profile. The spline baseline is constructed from cubic B-spline basis functions, including one additional knot outside either end of the user-specified fit range, as in LCModel. In contrast to LCModel, the baseline is not regularized.

Osprey follows the general workflow of LCModel. However, because the baseline curvature is not regularized, the baseline knot spacing is set to 0.15 ppm for preliminary modelling and increased to 0.4 ppm for the final full model.

Tarquin

The Tarquin (4.3.10) algorithm (Wilson et al., 2011) uses a four-step approach in the time domain to model the spectra. First, residual water is removed using singular value decomposition. Second, the global zero-order phase is determined by minimizing the difference between the magnitude and the real spectra in the frequency domain. Third, zero-filling to double the number of points and frequency referencing are performed, as in the other algorithms. This step also estimates a starting value for the Gaussian linebroadening used in the fourth step, the final modelling. The model includes common Gaussian linebroadening, individual Lorentzian linebroadening, individual frequency-shifts, and zero- and first-order phase correction factors applied in the frequency domain.

Optimization is performed in the time domain with a constrained non-linear least-squares Levenberg-Marquardt solver, allowing hard and soft constraints on the parameters. In addition, the range of time-domain datapoints is limited by removing the first 10 ms of the FID, so as to omit the fast-decaying macromolecule and lipid signals. Finally, the baseline is estimated in the frequency domain by convoluting the model residual with a Gaussian filter with a width of 100 points.

Model parameters

The parameters chosen for each tool are summarized in **Figure 1B**. The fit range was limited to 0.5 to 4 ppm in all tools to reduce effects of different water suppression techniques used across sites. For the baseline handling, the default parameters were chosen, i.e. blineKnotSpace = 0.4 ppm for Osprey, DKNMNT = 0.15 ppm for LCModel, and an FID range from 10 ms to 50% of

233 the FID for Tarquin. Note that for this diverse dataset, the 10-ms timepoint corresponded to a dif-
234 ferent number of points was defined in the Tarquin batch definition for different files.
235
236 Quantification, visualization, and secondary mechanistic analyses
237 The four major metabolite complexes tNAA (NAA + NAAG), tCho (GPC + PCh), mI, and Glx
238 (Glu + Gln) were quantified as basis-function amplitude ratios relative to total creatine (tCr = Cr
239 + PCr). Since the primary purpose was to compare performance of the core LCM algorithms, no
240 additional relaxation correction or partial volume correction was performed.
241
242 Model visualizations were generated with the *OspreyOverview* module, which allows LCModel
243 and Tarquin results files (.coord and .txt) to be imported. For each algorithm, the visualization
244 includes site-mean spectra, grand mean spectra (i.e. the mean of all spectra), and site- and grand
245 mean modelling results (complete model, spline baseline, spline baseline + MM components,
246 and the separate models of the major metabolite complexes). As for the default visualizations in
247 the LCModel and Tarquin software interfaces, inverse phase estimates are applied to the spectra
248 and final models. For the visualization, spectra were normalized to the amplitude of the 3-ppm
249 creatine singlet, and each site mean spectrum was vertically shifted to align the mean frequency-
250 domain amplitude between 1.85 and 4.0 ppm, to facilitate visual comparison between algorithms
251 and sites.
252
253 Two secondary mechanistic analyses were performed:
254 1. To investigate potential interactions between baseline power and metabolite estimates in a way
255 that was not strongly influenced by peak amplitudes, spectra were first aligned vertically
256 according to the frequency-domain minimum of the acquired data between 2.66 and 2.7 ppm (i.e.

between the aspartyl signals). Baseline power beneath each major metabolite was then defined as the range-normalized integral of the creatine-normalized baseline between 1.9 and 2.1 ppm for the tNAA baseline; 3.1 and 3.3 ppm for the tCho baseline; 3.33 and 3.75 ppm for mI; and 1.9 to 2.5 ppm and 3.6 to 3.8 ppm for the Glx baseline.

2. The contribution of variance in modelling of the creatine reference signal to metabolite ratios was also investigated. To this end, each individual total creatine model (Cr + PCr) was normalized to the frequency-domain amplitude of each metabolite spectrum between 1.9 and 2.1 ppm, and the integral over the individual creatine model was calculated.

Data analysis

Quantitative metabolite estimates (tNAA/tCr, tCho/tCr, mI/tCr, Glx/tCr) were statistically analyzed and visualized using R (R Core Team, 2017) in RStudio (Version 1.2.5019, RStudio Inc.). The functions are publicly available as a R-package in a GitHub repository (https://github.com/hezoe100/SpecVis, 2020). The supplemental materials with MATLAB- (job files and plot functions) and R-files (analysis script *LC_comp.R* and results *ResultsToPlot.RData*), example LCModel control files, and Tarquin batch-files are publicly available (Zöllner, 2020). The results from each LCM algorithm were imported into R with the *spvs_importResults* function using *read_tqn_result* and *read_lcm_coord* of the *spant* package (https://github.com/martin3141/spant, 2017) and R's standard *read.csv* function.

Distribution analysis

The results are presented as raincloud plots (Allen et al., 2019) and Pearson's correlation analysis using the *ggplot2* package (Wickham, 2009). The raincloud plots (*spvs_RainCloud*) include individual data points, boxplots with median and $25^{th}/75^{th}$ percentiles, a smoothed distribution, and mean ± SD error bars. In addition, the coefficient of variation (CV = SD/mean) and the mean

$\overline{CV} = \frac{(CV_{tNAA}+CV_{tCho}+CV_{Ins}+CV_{Glx})}{4}$ across all four metabolites of each algorithm (last column) are shown.

Correlation analysis

Global correlations (black), as well as within-vendor correlations (color-coded) are reported with different color shades for different sites (*spvs_Correlation* and *spvs_CorrelationFacet*). Furthermore, mean $\overline{\mathrm{R}^2}$ for each pair-wise correlation (e.g. Osprey vs LCModel) and metabolite, estimated by row or column means e.g. $\overline{R^2} = \frac{(R^2_{tNAA}+R^2_{tCho}+R^2_{Ins}+R^2_{Glx})}{4}$, and a grand mean $\overline{\mathbf{R}^2}$ (across all pair-wise correlations) are calculated. For the correlations, no correction for multiple testing was applied.

Statistical analysis

The global metabolite estimates were compared between algorithms with parametric tests, following recommendations for large sample sizes (Fagerland, 2012). Differences of variances were tested with Fligner-Killeen's test with a post-hoc pair-wise Fligner-Killeen's test and Bonferroni correction for the number of pair-wise comparisons. Depending on whether variances were different or not, an ANOVA or Welch's ANOVA was used to compare means with a post-hoc paired t-test with equal or non-equal variances, respectively (*spvs_Statistics*).

Results

All 277 spectra were successfully processed, exported, and quantified with the three LCM algorithms. After careful visual inspection of data and models for every algorithm by one author (HJZ, 6 years of MRS experience), no modelled spectra were excluded from further analysis.

Summary of the modelling results

A site-mean-level summary of the 277 spectra is shown in **Figure 2A**, **B** and **C**, for analyses in LCModel, Osprey, and Tarquin, respectively**.** Data, models and residuals are color-coded by vendor. A grand-mean summary separated by vendor is shown in Figure 2D< E and F (GE, Philips and Siemens, respectively). Data, models and residuals are color-coded by algorithm.

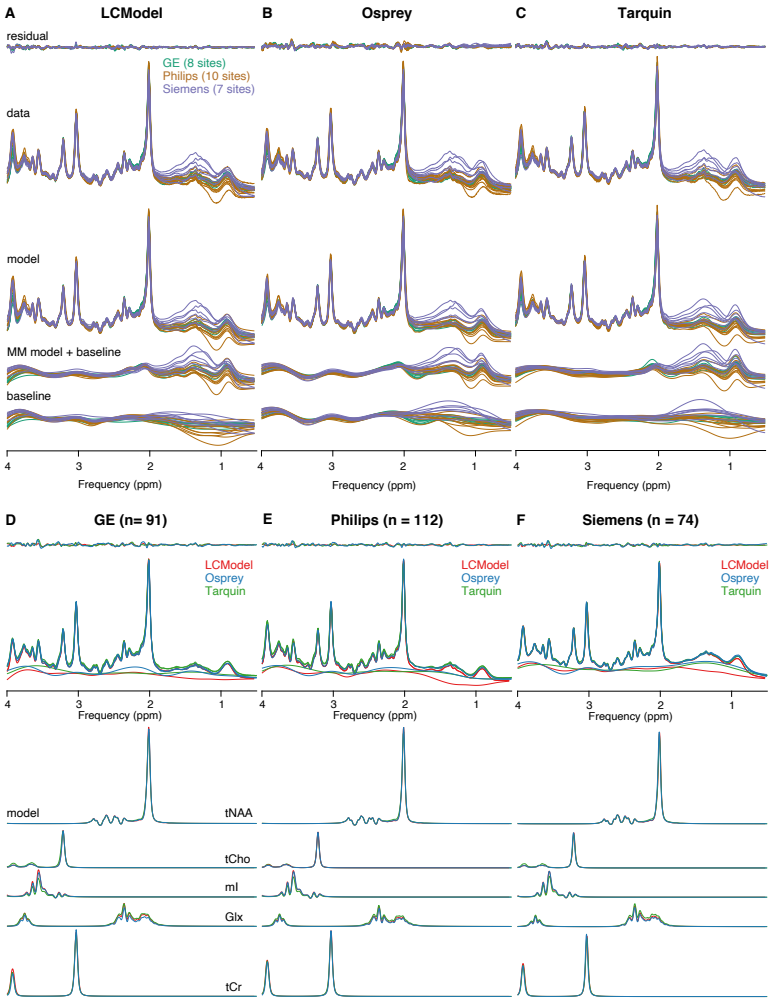


Figure 2- Summary of the modelling results. (A) – (C) site-mean residual, data, model, MM model + baseline, and baseline for each LCM algorithm, color-coded by vendor. (D) – (F) grand-mean residual, data, model, MM model + baseline, and metabolite models for each vendor, color-coded by LCM algorithm.

In general, the phased spectra and models agreed well between vendors for all algorithms. The greatest differences between vendors were found in the frequency ranges between 0.5 and 1.95 ppm, 2.8 and 2.95 ppm, and 3.68 and 4 ppm (as shown in **Figure 2A-C**), which were generally well-modelled by all algorithms. Structural features in the residual appear in the 1.9 to 2.1 ppm range for Osprey, between 2.2 and 2.6 ppm with a prominent underestimation at 2.28 ppm for all algorithms, and between 3.5 and 4 ppm for all algorithms with a prominent underestimation particularly for Osprey and Tarquin at 3.58 ppm. The amplitude of the residual is highest for Osprey, and similar for Tarquin and LCModel.

Comparing the algorithms, notable differences in the estimated spline baseline appeared between 0.5 and 1.95 ppm (degree of variability: Osprey > LCModel > Tarquin) and between 3.6 and 4 ppm (degree of variability: LCModel > Osprey > Tarquin). The baseline estimates were flattest for Tarquin, with the number of inflection points and the degree of curvature both following Osprey > LCModel > Tarquin.

The grand mean spectra and models agreed well across all vendors and algorithms (**Figure 2D-F**). Between-vendor agreement of the estimated baseline was higher between 2.6 and 3.1 ppm than elsewhere. The greatest baseline differences between algorithms occur between 0.5 and 1.95 ppm, with a higher agreement between Osprey and Tarquin compared to LCModel. A vendor-specific difference between algorithms appears above 3.6 ppm in the GE data **(Figure 2D)**.

Metabolite models agree well for tNAA, tCho, and tCr for all algorithms and vendors. The tCho resonances between 3.6 and 3.95 ppm (from GPC) were estimated higher by Tarquin for GE and

Siemens data than by the other algorithms. The Glx model appeared highest for Tarquin, followed by LCModel and Osprey, for all vendors. A very prominent difference between algorithms was the lower mI model in Tarquin.

Metabolite level distribution

*Table 2 – Metabolite level distribution. Mean, standard deviation and coefficient of variation (CV) of each metabolite-to-creatine ratio, listed by algorithm and vendor as well as global summary values. Asterisks indicate significant differences (adjusted p < 0.01 = ** or ## or ’’ and adjusted p < 0.001 = *** or ### or ’’’) in the mean (for the metabolite ratios) or the variance (for the CV) compared to the algorithm in the next row (LCModel vs Osprey = *, Osprey vs Tarquin = #, and Tarquin vs LCModel = ’).*

	[metabolite] / [tCr] (mean ± SD)			
	tNAA	tCho	mI	Glx
GE				
LCModel	1.48 ± 0.12	0.19 ± 0.02	0.85 ± 0.10	1.75 ± 0.25
Osprey	1.47 ± 0.10	0.18 ± 0.02	0.78 ± 0.09	1.42 ± 0.17
Tarquin	1.48 ± 0.11	0.22 ± 0.03	0.57 ± 0.07	2.05 ± 0.22
Philips				
LCModel	1.38 ± 0.10	0.17 ± 0.02	0.81 ± 0.08	1.46 ± 0.14
Osprey	1.50 ± 0.12	0.18 ± 0.02	0.86 ± 0.10	1.34 ± 0.16
Tarquin	1.40 ± 0.12	0.16 ± 0.03	0.60 ± 0.09	1.78 ± 0.19
Siemens				
LCModel	1.52 ± 0.19	0.19 ± 0.02	0.83 ± 0.09	1.65 ± 0.31
Osprey	1.54 ± 0.12	0.19 ± 0.02	0.89 ± 0.06	1.45 ± 0.14
Tarquin	1.50 ± 0.15	0.18 ± 0.03	0.65 ± 0.07	2.04 ± 0.19
global				
LCModel	1.45 ± 0.15***	0.18 ± 0.02	0.83 ± 0.09	1.45 ± 0.15***
Osprey	1.50 ± 0.12###	0.18 ± 0.02	0.84 ± 0.09###	1.50 ± 0.12###
Tarquin	1.46 ± 0.14	0.18 ± 0.04	0.60 ± 0.08’’’	1.93 ± 0.24’’’
	CV (SD/mean)			
	tNAA	tCho	mI	Glx
GE				
LCModel	7.9%	12.9%	11.8%	14.2%
Osprey	6.9%	9.7%	11.1%	11.8%
Tarquin	7.5%	11.7%	11.2%	10.8%
Philips				
LCModel	7.2%	10.6%	9.9%	9.7%
Osprey	8.0%	10.0%	11.8%	11.9%
Tarquin	8.8%	19.8%	15.2%	10.7%
Siemens				
LCModel	12.4%	13.4%	10.8%	18.7%
Osprey	8.0%	11.1%	6.9%	10.0%
Tarquin	10.1%	14.3%	10.5%	9.3%
global				
LCModel	10.0%	13.2%**	10.9%	16.4%***
Osprey	7.8%	10.4%###	11.7%###	11.8%###
Tarquin	9.3%	20.5%’’’	13.6%	12.3%

The tCr ratio estimates and CVs of the four metabolites are summarized in **Table 2**. Distributions and group statistics are visualized in **Figure 3**, with the four columns corresponding to the major metabolites, and the four rows corresponding the three vendors and a global summary across all datasets.

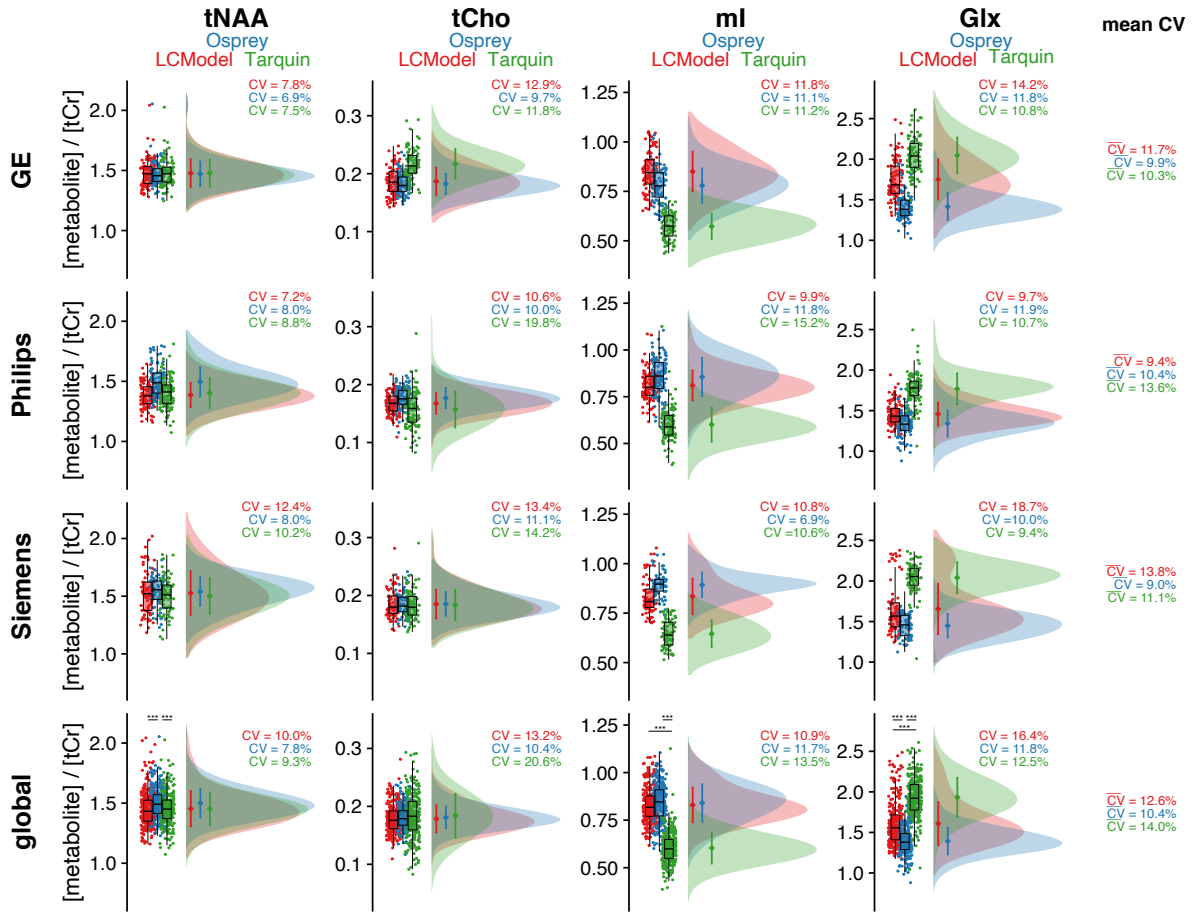


*Figure 3 – Metabolite level distribution. Raincloud plots of the metabolite estimates of each LCM algorithm (color-coded). The four metabolites are reported in the columns, and the three vendors in rows, with a global summary in the last row. The coefficient of variation is reported for each distribution, as well as a mean $\overline{CV}$ reported in the last column, which is calculated across each row. Asterisks indicate significant differences (adjusted $p < 0.001$ = ***).*

Generally speaking, between-algorithm agreement was greatest for the group means and CVs of tNAA and tCho. The grand-mean CV was lowest for Osprey (10.4%), followed by LCModel

(12.6%) and Tarquin (14.0%). Group means and CVs for tNAA are relatively consistent, although the mean tNAA from LCModel and Tarquin were slightly lower for Philips data than for Siemens and GE, while Osprey estimated tNAA similar for all vendors. As a result, the global tNAA/tCr was 1.45 ± 0.15 for LCModel, 1.50 ± 0.12 for Osprey, and 1.45 ± 0.14 for Tarquin, with significant differences between Osprey and both other LCM algorithms.

Group means and CVs for tCho showed the greatest agreement between Osprey and LCModel, while the estimates of Tarquin were notably higher for GE, and lower for Philips data. Consequently, the global CV of tCho estimates was significantly higher for Tarquin compared to both other algorithms, and significantly lower for Osprey compared to LCModel. Global tCho/tCr was 0.18 ± 0.02 for LCModel, 0.18 ± 0.02 for Osprey, and 0.18 ± 0.04 for Tarquin.

For mI, group means and CVs were comparable for Osprey and LCModel, while Tarquin estimates were lower by about 25%. The CV was notably smaller for Siemens spectra modelled in Osprey. The global CV was significantly lower for Osprey compared to Tarquin. Global mI/tCr was 0.83 ± 0.09 for LCModel, 0.84 ± 0.09 for Osprey, and 0.60 ± 0.08 for Tarquin, with significant mean differences between all algorithms.

Group means and CVs for Glx were comparable between Osprey and LCModel, while estimates were about 30% higher in Tarquin. The CV of LCModel estimates was higher for GE and Siemens data, with a bimodal distribution for the Siemens spectra. The global CV was significantly lower for Osprey compared to both other algorithms. The global Glx/tCr was 1.45 ± 0.15 for LCModel, 1.50 ± 0.12 for Osprey, and 1.93 ± 0.24 for Tarquin, with significant differences between all algorithms.

Mean $\overline{\text{CVs}}$, estimated by the row-mean, were between 9.0 and 13.8% for all algorithms and vendors.

Correlation analysis: pairwise comparison between LCM algorithms

The correlation analysis for each metabolite and algorithm pair is summarized in **Figure 4**. $\overline{R^2}$ for each algorithm pair and metabolite are reported in the corresponding row and column, respectively.

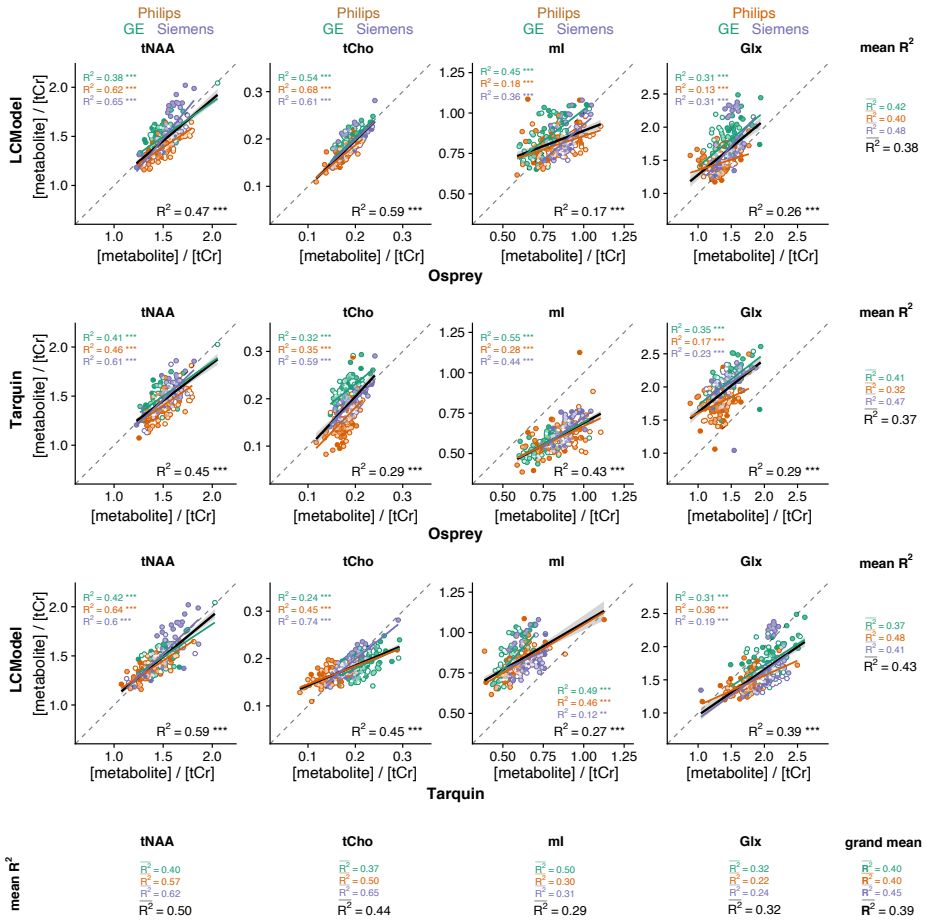


Figure 4 – Pairwise correlational comparison of algorithms. LCModel and Osprey are compared in the first row, Tarquin and Osprey are compared in the second row, and LCModel and Tarquin are compared in the third row. Each column corresponds to a different metabolite. Within-vendor correlations are color-coded and global correlations are shown in black. The $\overline{R^2}$ *values are calculated along each dimension of the grid with mean* R^2 *for each metabolite and each correlation. A grand mean* $\overline{\boldsymbol{R}^2}$ *value is also calculated across all twelve pair-wise correlations. Asterisks indicate significant correlations (p < 0.01 = ** and p < 0.001 = ***).*

The global grand mean $\overline{\mathbf{R}^2}$ = 0.39 suggests an overall moderate agreement between metabolite estimates from different algorithms. The agreement between algorithms, estimated by the row-mean $\overline{R^2}$, was highest for Tarquin-vs-LCModel ($\overline{R^2}$ = 0.43), followed by Osprey-vs-LCModel ($\overline{R^2}$ = 0.38), and Osprey-vs-Tarquin ($\overline{R^2}$ = 0.37).

The agreement between algorithm for each metabolite , estimated by the column-mean $\overline{R^2}$, was highest for tNAA ($\overline{R^2}$ = 0.50), followed by tCho ($\overline{R^2}$ = 0.44), Glx ($\overline{R^2}$ = 0.32), and mI ($\overline{R^2}$ = 0.29). The grand mean $\overline{\mathbf{R}^2}$ for each vendor was higher for Siemens ($\overline{\mathbf{R}^2}$ = 0.45) than for GE ($\overline{\mathbf{R}^2}$ = 0.40) and Philips ($\overline{\mathbf{R}^2}$ = 0.40).

While the within-metabolite mean $\overline{R^2}$ (average down the columns in Figure 4) are comparable between vendors, there is substantially higher variability of the R^2 values with increasing granularity of the analysis. As an example, Siemens tCho estimates agree exceptionally well between Tarquin and LCModel (R^2 = 0.74) while Siemens mI estimates of Tarquin vs LCModel agree very poorly (R^2 = 0.12). **Inline Supplementary Material 1** includes an additional layer of correlations at the site level. These correlations were found to be even more inconsistent than correlations at global or within-vendor levels, emphasizing the caution that should be applied in considering metabolite correlations in small sample sizes.

<u>Correlation analysis: baseline and metabolite estimates</u>

The correlation analysis between local baseline power and metabolite estimates for each algorithm is summarized in **Figure 5**. The global grand mean $\overline{\mathbf{R}^2}$ = 0.10 suggests that overall, there is

an association between local baseline power and metabolite estimates, that is weak but statistically significant. The influence of baseline on metabolite estimates differs between metabolites, as reflected by the column-mean $\overline{R^2}$ which was lowest for tCho ($\overline{R^2}$ = 0.04) and tNAA ($\overline{R^2}$ = 0.06), and higher for mI ($\overline{R^2}$ = 0.13) and Glx ($\overline{R^2}$ = 0.18). Additionally, the direction of the global correlation was found to be negative for all correlations, except for tCho estimates of Tarquin.

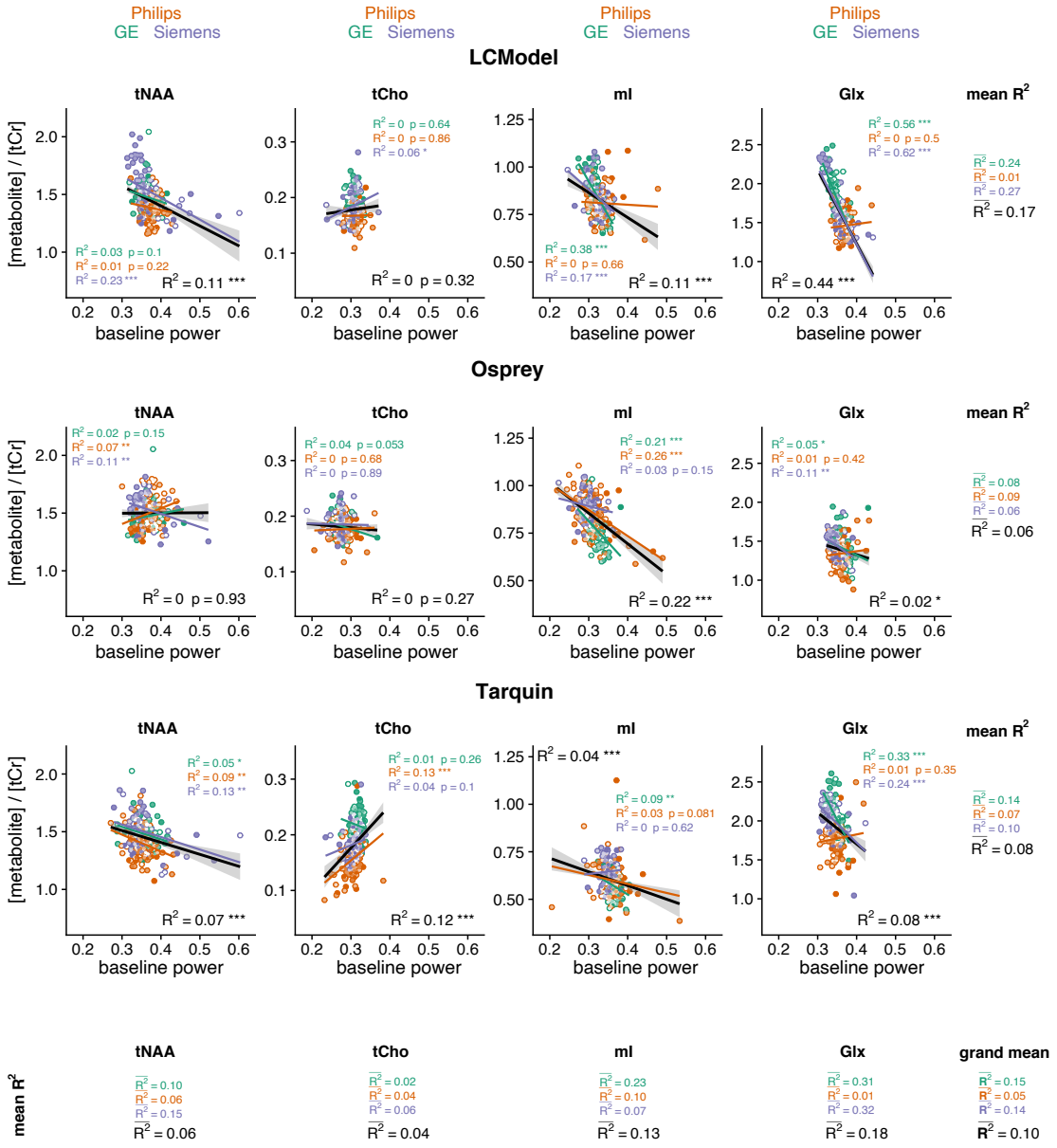


*Figure 5– Correlation analysis between metabolite estimates and local baseline power for each algorithm, including global (black) and within-vendor (color-coded) correlations. The mean $\overline{R^2}$ values are calculated along each dimension of the grid for each metabolite and each algorithm. Similarly, a grand mean $\overline{\boldsymbol{R}^2}$ value is calculated across all twelve pair-wise correlations. Asterisks indicate significant correlations ($p < 0.05$ = *, $p < 0.01$ = ** ,and $p < 0.001$ = ***).*

The mean $\overline{R^2}$ across metabolites for each algorithm, calculated as the row mean, were low for all algorithms with LCModel ($\overline{R^2}$ = 0.17) showing a greater effect than Tarquin ($\overline{R^2}$ = 0.08) and Osprey ($\overline{R^2}$ = 0.06). Comparing between vendors, the grand-mean $\overline{\mathbf{R}^2}$ was higher for GE ($\overline{\mathbf{R}^2}$ = 0.15) and Siemens ($\overline{\mathbf{R}^2}$ = 0.14) than for Philips ($\overline{\mathbf{R}^2}$ = 0.05) spectra.

Relationships between metabolite estimates and local baseline power were extremely diverse at a finer level of analysis. As an example, R^2 varied from 0.62 for Siemens Glx in LCModel to 0.00 for Siemens and Philips tCho in Osprey.

Variability of total creatine models

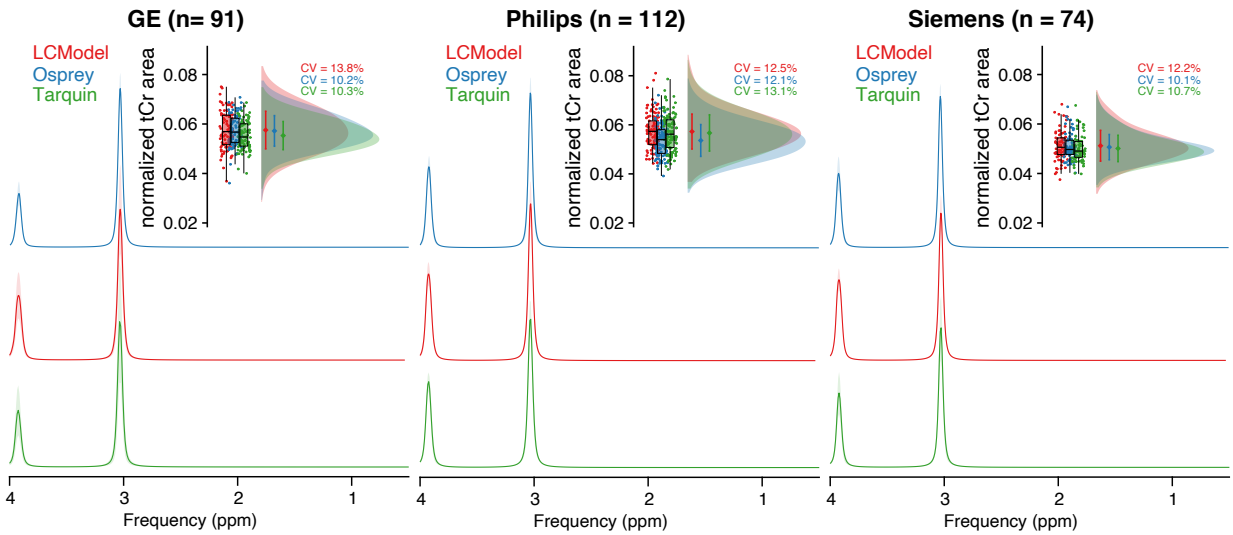


Figure 6 – Variability of tCr models. Mean models +/- standard deviation (shaded areas) are presented column-wise by vendor and color-coded by LCM algorithm. The distribution and CV of the areas under the models are inset

Mean tCr model spectra (plus-minus one standard deviations) are summarized in **Figure 6** for each vendor and LCM algorithm, along with distribution plots of the area under the model . The mean GE models indicate a higher estimate of the negative -$CrCH_2$ resonance for all LCM compared to the other vendors, reflected in a smaller amplitude of these resonances, most likely a result of broader water suppression on the GE systems. The agreement in mean and CV is greatest between Osprey and Tarquin for all vendors, while tCr areas for LCModel appear slightly higher.

Discussion

We have presented a three-way comparison study of LCM algorithms applied to a large dataset of short-TE in-vivo human brain MRS spectra. The aims at the onset were to compare metabolite estimates obtained with different LCM algorithms, and to identify potential sources of differences between the algorithms. The major findings are:

- Group means and CVs for tNAA and tCho agreed well across vendors and algorithms. For mI and Glx, group means and CVs were less consistent between algorithms, with a higher degree of agreement between Osprey and LCModel compared to Tarquin.
- Despite the large number of datasets included, the strength of the correlations between individual metabolite estimates from different algorithms was moderate. In general, tNAA and tCho estimates from different algorithms agreed better than Glx and mI. With each sub-level of analysis, the variability of correlation strength increased, i.e. correlations grew increasingly variable when calculated separately for each vendor, or even each site.
- Overall, the association between metabolite estimates and the local baseline power was significant, with mI and Glx showing stronger associations than tNAA and tCho, and LCModel showing greater effects than Tarquin and Osprey.

The strong agreement of group means and CVs for metabolites with prominent singlets (tNAA/tCho) and greater inconsistency for lower-intensity coupled signals (mI/Glx) are in line with previous two-tool comparisons of simulated data (Mosconi et al., 2014; Wilson et al., 2011) and in-vivo studies with smaller sample sizes (Kossowski et al., 2017; Scott et al., 2016; Wilson et al., 2011).

While previous work highlighted group means and standard deviations, the between-algorithm agreement of individual metabolite estimates has not been extensively studied. Our results suggest that substantial variability is introduced by the choice of the analysis software itself, indicated by only moderate between-algorithm correlation strength (between-algorithm mean $\overline{R^2}$ <= 0.5 for all investigated metabolites), even for the well-established LCM algorithms LCModel and Tarquin (R^2 between 0.27 and 0.59 for all metabolites). This finding raises concerns about the robustness of correlative MRS study results, as the additional variability arising from the choice of analysis tool is rarely recognized or acknowledged. MRS studies typically suffer from low sample sizes, and with this level of algorithm-dependent variance, it is likely that marginally significant group effects and correlations found with one analysis tool at one site might not be found with another. This is evident from the substantial variability of correlation strengths at vendor- or even site-level, and is even more likely to be the case for 'real-life' clinical data, given the relatively high quality of the dataset in this study (standardized pre-processing; large sample size; high SNR; low linewidth; young, healthy, cooperative subjects). While two previous studies found that some differences between clinical groups remained significant independent of the LCM algorithm (Kossowski et al., 2017; Scott et al., 2016), it is questionable to make this a default assumption. If choice of analysis tool is a significant contributor to measurement variance, it could be argued that modelling of data with more than one algorithm will improve the robustness and power of MRS studies.

Sources of variance

In order to understand the substantial variability introduced by the choice of analysis tool, the influence of modelling strategies and parameters on quantitative results needs to be better understood. Previous investigations have shown that, within a given LCM algorithm, metabolite

estimates can be affected by the choice of baseline knot spacing (Marjańska and Terpstra, 2019; Wenger et al., 2019), the modelling of MM and lipids (Marjańska and Terpstra, 2019; Schaller et al., 2014), and SNR and linewidth (Bartha, 2005; Near, 2014; Wijtenburg and Knight-Scott, 2009; Zhang and Shen, 2020). In this study, we focused on the comparison of each LCM with their default parameters, and observed differences result both from these parameters and from differences in the core algorithm.

The process of LCM relies on the assumption that broad background and baseline signals can be separated from narrower metabolite signals. This is true to a limited degree, and the choice of MM and baseline modelling influences the quantification of metabolite resonances (Near et al., 2020). As previously shown for the same dataset used in this study (Považan et al., 2020), metabolite estimates are relatively consistent between vendors, more so than between algorithms. This further supports the notion that MM/baseline modelling is a critical source of variance in metabolite estimates. While a recent consensus paper recommends including MM basis functions in modelling, there is no consensus on how to achieve this – MM basis functions are currently not consistently defined and handled. Baseline modelling is even more diversely implemented across algorithms. Our secondary analysis of the relationship between baseline power and metabolite estimates showed a stronger interaction for the broader coupled signals of Glx and mI than the singlets. Cho showed the weakest effect, likely due to its proximity in the spectrum to the Cr reference signal. These findings suggest that the higher variance of Glx and mI estimates are at least partly explained by stronger coupling to the local baseline power. This would imply that separability of metabolite and macromolecule signals is insufficient in current modelling approaches, emphasizing the need for advanced MM + baseline modelling strategies (Marjańska and Terpstra, 2019).

While including experimental MM acquisitions into the study may reduce the degrees of freedom of the modelling process itself, other sources of variance, such as age-dependency (Marjańska et al., 2018) or tissue composition (Považan et al., 2018; Schaller et al., 2014) remain, and there are many open questions on how to address this issue (Cudalbu et al., 2020).

Since this study focused on reporting tCr ratios, it is important to consider the variance of the creatine model of each algorithm. With MRS only quantitative in a relative sense, separating the variance contribution of the reference signal is a challenge. While mean tCr model areas were slightly higher for LCModel than for Osprey and Tarquin, there was no generalizable observation of lower tCr ratios from LCModel. CVs of the tCr model areas were comparable across LCM algorithms for each vendor. Taken together, the contribution of the creatine reference signal to the variance of metabolite estimates is unclear and hard to isolate.

Limitations

As mentioned above, there is currently no widely adopted consensus on the definition of MM basis functions. Therefore, the default LCModel MM basis function definitions from LCModel (section 11.7 of the LCModel manual) were adapted for each algorithm. Additionally, these basis functions only included MMs for frequencies < 3.0 ppm, which is likely insufficient for the modelling of MM signals between 3 and 4 ppm (Giapitzakis et al., 2018), and will have repercussions for the estimation of tCho, mI, and Glx. Second, standard modelling parameters were chosen for each LCM, which ensure a broader comparability to the current literature, but may provide suboptimal results – e.g. recent recommendations call for the LCModel DKNTMN parameter to be set to >=0.25, resulting in less flexible baselines. Third, there is obviously no ‘gold standard’

508 of metabolite level estimation to validate MRS results against. The performance of an algorithm
509 is often judged based on the level of variance, but low variance clearly does not reflect accuracy
510 and may indicate insufficient responsiveness of a model to the data. In comparing multiple algo-
511 rithms, it is tempting to infer algorithms that show a higher degree of correlation in results are
512 more reliable, but it could equally be the case that shared algorithm-based sources of variance
513 increase such correlations. Efforts to use simulated spectra as a gold-standard, including those
514 applying machine learning (Lee and Kim, 2020, 2019), can only be successful to the extent that
515 simulated data are truly representative of in-vivo data. Fourth, this study was limited to the two
516 most widely used algorithms LCModel and Tarquin, as well as the Osprey algorithm that is un-
517 der ongoing development in our group.
518 While including additional algorithms would increase the general understanding of different al-
519 gorithms, the complexity of the resulting analysis and interpretation would be overwhelming and
520 beyond the scope of a single publication.
521

522 Future developments

523 This study emphasizes the need to further investigate LCM algorithms to understand the
524 influence of various factors on linear-combination modelling of MRS data. It showcases a
525 standardized framework to investigate the performance of LCM algorithms and to visualize the
526 outcome, using large-scale datasets, our new open-source software Osprey, and the developed R
527 software (https://github.com/hezoe100/SpecVis, 2020) (future releases should also include
528 import functions for results from other LCM tools). Several avenues of investigation can be
529 conveniently explored within this framework including those that are accessible as parameters in
530 compiled tools such as LCModel and hidden within the algorithm. Firstly, the sensitivity of

LCM algorithms could be tested to small known perturbations of the spectrum, e.g. with Monte-Carlo-like modelling techniques introducing small changes to the model and re-fitting. Secondly, different baseline models could be compared by incorporating acquired MM data, modifying the number and parametrization of MM basis functions, or testing different baseline modelling approaches (impact of knot spacing, regularization, or an adaptative baseline model (Wilson, 2020)). Thirdly, the impact of modifying crucial LCM parameters could be investigated such as hard and soft constraints, fit range in the frequency domain or the number of points included in the time domain fit, or manipulating starting values for non-linear parameters. Fourthly, the impact of changes to pre-processing (and preliminary modelling), e.g. by testing different SNR improvement approaches, or modifying the routines used to generate starting values for the modelling could be studied. Fifthly, similar to the training of neural network algorithms, noisy simulated data may be used to create a ground truth to determine modelling precision and accuracy.

Conclusion

This study presents a comparison of three LCM algorithms applied to a large short-TE PRESS dataset. The open-source software tools applied represent a framework for future studies of the LCM process itself.

While different linear-combination modelling algorithms' estimates of major metabolite levels may broadly agree at a group level, correlations between results are only weak-to-moderate, despite standardized pre-processing, a large sample of young, healthy and cooperative subjects, and high spectral quality. The variability of metabolite estimates that is introduced by the choice of analysis software is substantial, raising concerns about the robustness of MRS research findings, which typically use a single algorithm to draw inferences from much smaller sample sizes.

References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Kievit, R.A., 2019. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res. 4, 63. https://doi.org/10.12688/wellcomeopenres.15191.1

Barkhuijsen, H., de Beer, R., van Ormondt, D., 1987. Improved algorithm for noniterative time-domain model fitting to exponentially damped magnetic resonance signals. J. Magn. Reson. 1969 73, 553–557. https://doi.org/10.1016/0022-2364(87)90023-0

Bartha, R., 2005. The Effect of Signal to Noise Ratio and Linewidth On 4T Short Echo Time 1H MRS Metabolite Quantification. Proc. 13th Sci. Meet. Int. Soc. Magn. Reson. Med. 216, 2459–2459.

Bottomley, P., 1985. Selective volume method for performing localized NMR spectroscopy, United States Patent and Trademark Office. https://doi.org/10.1016/0730-725X(85)90032-3

Cudalbu, C., Behar, K.L., Bhattacharyya, P.K., Bogner, W., Borbath, T., de Graaf, R.A., Henning, A., Juchem, C., Kreis, R., Lee, P., Hongxia, L., Marjańska, M., Mekle, R., Murali-Manohar, S., Považan, M., Rackayová, V., Wright, A.M., Simicic, D., Slotboom, J., Starcuk, Z., Starčuková, J., Soher, B.J., Tkáč, I., Williams, S., Wilson, M., Xin, L., Mlynrik, V., 2020. Contribution of macromolecules to brain 1H MR spectra: Experts' consensus recommendations. NMR Biomed. Revis.

Cudalbu, C., Cavassila, S., Rabeson, H., van Ormondt, D., Graveron-Demilly, D., 2008. Influence of measured and simulated basis sets on metabolite concentration estimates. NMR Biomed. 21, 627–636. https://doi.org/10.1002/nbm.1234

Fagerland, M.W., 2012. T-tests, non-parametric tests, and large studiesa paradox of statistical practice? BMC Med. Res. Methodol. 12, 78. https://doi.org/10.1186/1471-2288-12-78

Giapitzakis, I.-A., Avdievich, N., Henning, A., 2018. Characterization of macromolecular baseline of human brain using metabolite cycled semi-LASER at 9.4T. Magn. Reson. Med. 80, 462–473. https://doi.org/10.1002/mrm.27070

Gottschalk, M., Lamalle, L., Segebarth, C., 2008. Short-TE Iocalised 1H MRS of the human brain at 3T: Quantification of the metabolite signals using two approaches to account for macromolecular signal contributions. NMR Biomed. 21, 507–517. https://doi.org/10.1002/nbm.1219

Graveron-Demilly, D., 2014. Quantification in magnetic resonance spectroscopy based on semi-parametric approaches. Magn. Reson. Mater. Phys. Biol. Med. 27, 113–130. https://doi.org/10.1007/s10334-013-0393-4

Hall, E.L., Stephenson, M.C., Price, D., Morris, P.G., 2014. Methodology for improved detection of low concentration metabolites in MRS: Optimised combination of signals from multi-element coil arrays. NeuroImage 86, 35–42. https://doi.org/10.1016/j.neuroimage.2013.04.077

https://github.com/hezoe100/SpecVis [WWW Document], 2020. . SpecVis GitHub Repos. URL https://github.com/hezoe100/SpecVis (accessed 5.27.20).

https://github.com/martin3141/spant [WWW Document], 2017. . Spant GitHub Repos. URL https://github.com/martin3141/spant (accessed 5.27.20).

https://github.com/schorschinho/osprey [WWW Document], 2020. . Osprey GitHub Repos. URL https://github.com/schorschinho/osprey (accessed 5.27.20).

https://www.nitrc.org/projects/biggaba/ [WWW Document], 2018. . Big GABA Repos. URL https://www.nitrc.org/projects/biggaba/ (accessed 5.27.20).

Juchem, C., 2018. INSPECTOR - A Tool for Teaching Magnetic Resonance Spectroscopy, in: 26th Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM). Paris, France.

Klose, U., 1990. In vivo proton spectroscopy in presence of eddy currents. Magn. Reson. Med. 14, 26–30. https://doi.org/10.1002/mrm.1910140104

Kossowski, B., Orzeł, J., Bogorodzki, P., Wilson, M., Setkowicz, Z., P. Gazdzinski, S., 2017. Follow-up analyses on the effects of long-term use of high fat diet on hippocampal metabolite concentrations in Wistar rats: Comparing Tarquin quantification of 7.0T rat metabolites to LCModel. Biol. Eng. Med. 2. https://doi.org/10.15761/BEM.1000129

Lee, H.H., Kim, H., 2020. Deep learning-based target metabolite isolation and big data-driven measurement uncertainty estimation in proton magnetic resonance spectroscopy of the brain. Magn. Reson. Med. n/a. https://doi.org/10.1002/mrm.28234

Lee, H.H., Kim, H., 2019. Intact metabolite spectrum mining by deep learning in proton magnetic resonance spectroscopy of the brain. Magn. Reson. Med. 82, 33–48. https://doi.org/10.1002/mrm.27727

Levenberg, K., 1944. A method for the solution of certain non-linear problems in least squares. Q. Appl. Math. 2, 164–168. https://doi.org/10.1090/qam/10666

Marjańska, M., Deelchand, D.K., Hodges, J.S., McCarten, J.R., Hemmy, L.S., Grant, A., Terpstra, M., 2018. Altered macromolecular pattern and content in the aging human brain. NMR Biomed. 31, e3865. https://doi.org/10.1002/nbm.3865

Marjańska, M., Terpstra, M., 2019. Influence of fitting approaches in LCModel on MRS quantification focusing on age-specific macromolecules and the spline baseline. NMR Biomed. https://doi.org/10.1002/nbm.4197

Marquardt, D.W., 1963. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. J. Soc. Ind. Appl. Math. 11, 431–441. https://doi.org/10.1137/0111030

Mikkelsen, M., Near, J., Saleh, M.G., Mostofsky, S.H., Puts, N.A.J., Edden, R.A.E., 2019. Robust Correction of Frequency and Phase Errors in Edited MRS Data, in: 28th Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM).

Mosconi, E., Sima, D.M., Garcia, M.I.O., Fontanella, M., Fiorini, S., Huffel, S.V., Marzola, P., 2014. Different quantification algorithms may lead to different results: a comparison using proton MRS lipid signals. NMR Biomed. 27, 431–443. https://doi.org/10.1002/nbm.3079

Naressi, A., Couturier, C., Devos, J.M., Janssen, M., Mangeat, C., Beer, R. de, Graveron-Demilly, D., 2001. Java-based graphical user interface for the MRUI quantitation package. Magn. Reson. Mater. Phys. Biol. Med. 12, 141. https://doi.org/10.1007/BF02668096

Near, J., 2014. Investigating the effect of spectral linewidth on metabolite measurement bias in short-TE MRS, in: 21th Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM). Milan, Italy.

Near, J., Harris, A.D., Juchem, C., Kreis, R., Marjańska, M., Öz, G., Slotboom, J., Wilson, M., Gasparovic, C., 2020. Preprocessing, analysis and quantification in single-voxel magnetic resonance spectroscopy: experts' consensus recommendations. NMR Biomed. n/a, e4257. https://doi.org/10.1002/nbm.4257

Oeltzschner, G., Zöllner, H.J., Hui, S.C.N., Mikkelsen, M., Saleh, M.G., Tapper, S., Edden, R.A.E., 2020. Osprey: Open-Source Processing, Reconstruction & Estimation of Magnetic Resonance Spectroscopy Data. bioRxiv. https://doi.org/10.1101/2020.02.12.944207

Osorio-Garcia, M.I., Sima, D.M., Nielsen, F.U., Himmelreich, U., Huffel, S.V., 2011. Quantification of magnetic resonance spectroscopy signals with lineshape estimation. J. Chemom. 25, 183–192. https://doi.org/10.1002/cem.1353

Öz, G., Alger, J.R., Barker, P.B., Bartha, R., Bizzi, A., Boesch, C., Bolan, P.J., Brindle, K.M., Cudalbu, C., Dinçer, A., Dydak, U., Emir, U.E., Frahm, J., González, R.G., Gruber, S., Gruetter, R., Gupta, R.K., Heerschap, A., Henning, A., Hetherington, H.P., Howe, F.A., Hüppi, P.S., Hurd, R.E., Kantarci, K., Klomp, D.W.J., Kreis, R., Kruiskamp, M.J., Leach, M.O., Lin, A.P., Luijten, P.R., Marjańska, M., Maudsley, A.A., Meyerhoff, D.J., Mountford, C.E., Nelson, S.J., Pamir, M.N., Pan, J.W., Peet, A.C., Poptani, H., Posse, S., Pouwels, P.J.W., Ratai, E.-M., Ross, B.D., Scheenen, T.W.J., Schuster, C., Smith, I.C.P., Soher, B.J., Tkáč, I., Vigneron, D.B., Kauppinen, R.A., 2014. Clinical Proton MR Spectroscopy in Central Nervous System Disorders. Radiology 270, 658–679. https://doi.org/10.1148/radiol.13130531

Poullet, J.-B., Sima, D.M., Simonetti, A.W., De Neuter, B., Vanhamme, L., Lemmerling, P., Van Huffel, S., 2007. An automated quantitation of short echo time MRS spectra in an open source software environment: AQSES. NMR Biomed. 20, 493–504. https://doi.org/10.1002/nbm.1112

Považan, M., Mikkelsen, M., Berrington, A., Bhattacharyya, P.K., Brix, M.K., Buur, P.F., Cecil, K.M., Chan, K.L., Chen, D.Y.T., Craven, A.R., Cuypers, K., Dacko, M., Duncan, N.W., Dydak, U., Edmondson, D.A., Ende, G., Ersland, L., Forbes, M.A., Gao, F., Greenhouse, I., Harris, A.D., He, N., Heba, S., Hoggard, N., Hsu, T.-W., Jansen, J.F.A., Kangarlu, A., Lange, T., Lebel, R.M., Li, Y., Lin, C.-Y.E., Liou, J.-K., Lirng, J.-F., Liu, F., Long, J.R., Ma, R., Maes, C., Moreno-Ortega, M., Murray, S.O., Noah, S., Noeske, R., Noseworthy, M.D., Oeltzschner, G., Porges, E.C., Prisciandaro, J.J., Puts, N.A.J., Roberts, T.P.L., Sack, M., Sailasuta, N., Saleh, M.G., Schallmo, M.-P., Simard, N., Stoffers, D., Swinnen, S.P., Tegenthoff, M., Truong, P., Wang, G., Wilkinson, I.D., Wittsack, H.-J., Woods, A.J., Xu, H., Yan, F., Zhang, C., Zipunnikov, V., Zöllner, H.J., Edden, R.A.E., Barker, P.B., 2020. Comparison of Multivendor Single-Voxel MR Spectroscopy Data Acquired in Healthy Brain at 26 Sites. Radiology 295, 191037. https://doi.org/10.1148/radiol.2020191037

Považan, M., Strasser, B., Hangel, G., Heckova, E., Gruber, S., Trattnig, S., Bogner, W., 2018. Simultaneous mapping of metabolites and individual macromolecular components via ultra-short acquisition delay 1H MRSI in the brain at 7T. Magn. Reson. Med. 79, 1231–1240. https://doi.org/10.1002/mrm.26778

Provencher, S.W., 1993. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. Magn. Reson. Med. 30, 672–679. https://doi.org/10.1002/mrm.1910300604

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schaller, B., Xin, L., Gruetter, R., 2014. Is the macromolecule signal tissue-specific in healthy human brain? A \textlesssup\textgreater1\textless/sup\textgreater H MRS study at 7 tesla in the occipital lobe. Magn. Reson. Med. 72, 934–940. https://doi.org/10.1002/mrm.24995

Scott, J., Underwood, J., Garvey, L.J., Mora-Peris, B., Winston, A., 2016. A comparison of two post-processing analysis methods to quantify cerebral metabolites measured via proton

magnetic resonance spectroscopy in HIV disease. Br. J. Radiol. 89, 20150979. https://doi.org/10.1259/bjr.20150979

Shen, Z.W., Chen, Y.W., Wang, H.Y., Lin, Y., He, J.K., Liu, W.W., Chen, F.Y., Wu, R.H., 2008. Quantification of Metabolites in Swine Brain by ^1H MR Spectroscopy Using LCModel and QUEST: A Comparison Study, in: 2008 Congress on Image and Signal Processing. Presented at the 2008 Congress on Image and Signal Processing, pp. 299–302. https://doi.org/10.1109/CISP.2008.478

Simpson, R., Devenyi, G.A., Jezzard, P., Hennessy, T.J., Near, J., 2017. Advanced processing and simulation of MRS data using the FID appliance (FID-A)—An open source, MATLAB-based toolkit. Magn. Reson. Med. 77, 23–33. https://doi.org/10.1002/mrm.26091

Soher, B.J., Semanchuk, P., Todd, D., Steinberg, J., Young, K., 2011. VeSPA: Integrated applications for RF pulse design, spectral simulation and MRS data analysis, in: 19th Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM). Montreal, Canada.

Stefan, D., Cesare, F.D., Andrasescu, A., Popa, E., Lazariev, A., Vescovo, E., Strbak, O., Williams, S., Starcuk, Z., Cabanas, M., van Ormondt, D., Graveron-Demilly, D., 2009. Quantitation of magnetic resonance spectroscopy signals: the jMRUI software package. Meas. Sci. Technol. 20, 104035. https://doi.org/10.1088/0957-0233/20/10/104035

Vanhamme, L., van den Boogaart, A., Van Huffel, S., 1997. Improved Method for Accurate and Efficient Quantification of MRS Data with Use of Prior Knowledge. J. Magn. Reson. 129, 35–43. https://doi.org/10.1006/JMRE.1997.1244

Wenger, K.J., Hattingen, E., Harter, P.N., Richter, C., Franz, K., Steinbach, J.P., Bähr, O., Pilatus, U., 2019. Fitting algorithms and baseline correction influence the results of non-invasive in vivo quantitation of 2-hydroxyglutarate with 1H-MRS. NMR Biomed. 32, e4027. https://doi.org/10.1002/nbm.4027

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wijtenburg, S.A., Knight-Scott, J., 2009. The Impact of SNR on the Reliability of LCModel and QUEST Quantitation in 1 H-MRS, in: 17th Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM).

Wilson, M., 2020. Adaptive Baseline Fitting for 1H MR Spectroscopy Analysis. bioRxiv 2020.02.17.949495. https://doi.org/10.1101/2020.02.17.949495

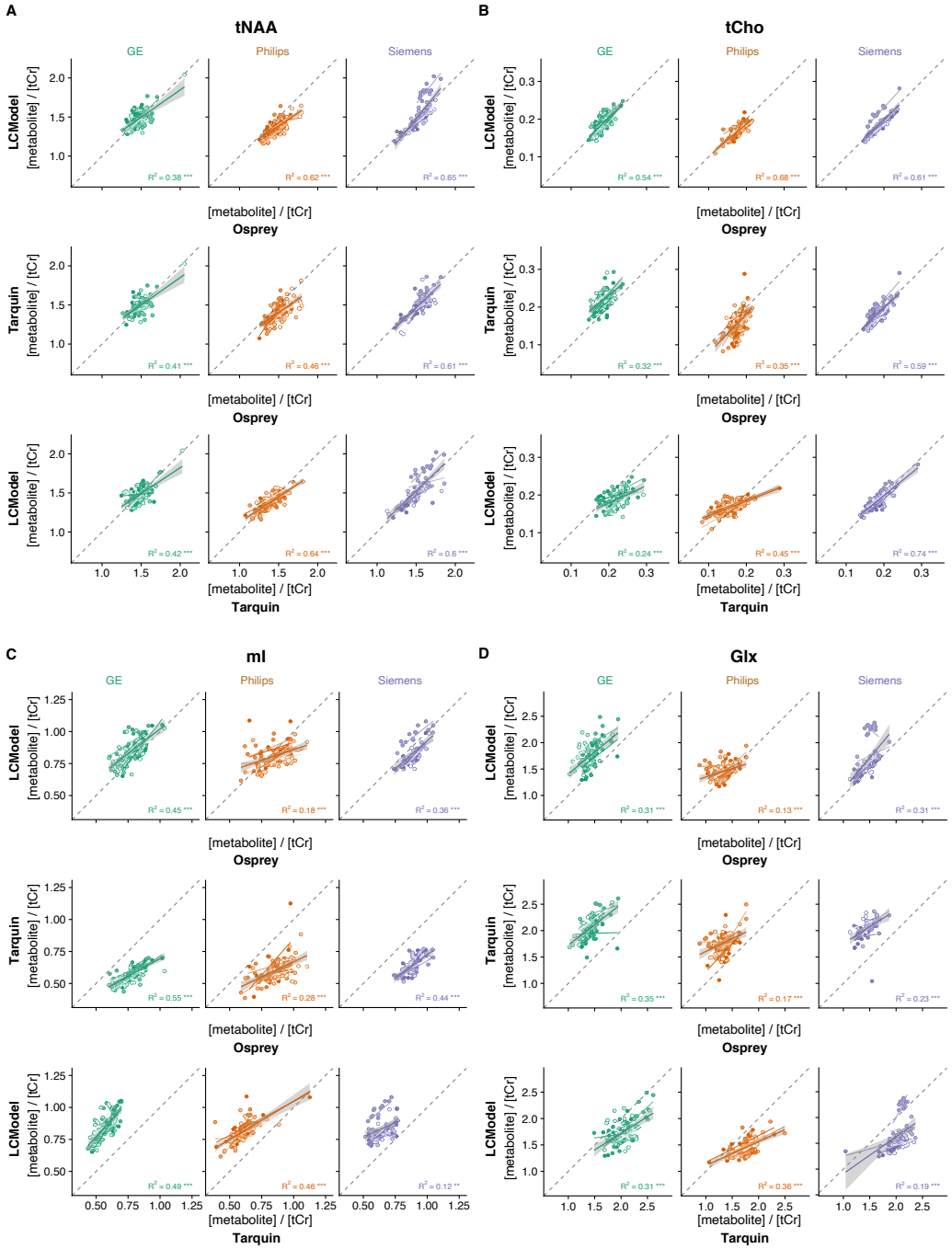
Wilson, M., Andronesi, O., Barker, P.B., Bartha, R., Bizzi, A., Bolan, P.J., Brindle, K.M., Choi, I., Cudalbu, C., Dydak, U., Emir, U.E., Gonzalez, R.G., Gruber, S., Gruetter, R., Gupta, R.K., Heerschap, A., Henning, A., Hetherington, H.P., Huppi, P.S., Hurd, R.E., Kantarci, K., Kauppinen, R.A., Klomp, D.W.J., Kreis, R., Kruiskamp, M.J., Leach, M.O., Lin, A.P., Luijten, P.R., Marjańska, M., Maudsley, A.A., Meyerhoff, D.J., Mountford, C.E., Mullins, P.G., Murdoch, J.B., Nelson, S.J., Noeske, R., Öz, G., Pan, J.W., Peet, A.C., Poptani, H., Posse, S., Ratai, E., Salibi, N., Scheenen, T.W.J., Smith, I.C.P., Soher, B.J., Tkáč, I., Vigneron, D.B., Howe, F.A., 2019. Methodological consensus on clinical proton MRS of the brain: Review and recommendations. Magn. Reson. Med. 82, 527–550. https://doi.org/10.1002/mrm.27742

Wilson, M., Reynolds, G., Kauppinen, R.A., Arvanitis, T.N., Peet, A.C., 2011. A constrained least-squares approach to the automated quantitation of in vivo 1 H magnetic resonance spectroscopy data. Magn. Reson. Med. 65, 1–12. https://doi.org/10.1002/mrm.22579

Zhang, Y., Shen, J., 2020. Effects of noise and linewidth on in vivo analysis of glutamate at 3 T. J. Magn. Reson. 314. https://doi.org/10.1016/j.jmr.2020.106732

Zöllner, H.J., 2020. Comparison of algorithms for linear-combination modelling of short-echo-time magnetic resonance spectra [WWW Document]. URL https://osf.io/3ekq4/ (accessed 6.2.20).

Inline Supplementary Material



*Inline Supplementary Material 1 – Facetted pair-wise correlational comparison of algorithms. LCModel and Osprey are compared in the first row, Tarquin and Osprey are compared in the second row, and LCModel and Tarquin are compared in the third row. Each sub-plot (A-D) corresponds to a different metabolite. Within-vendor (bold line with confidence interval) and within-site (thin line) correlations are color-coded. Asterisks indicate significant correlations (p < 0.01 = ** and p < 0.001 = ***).*

Declaration of competing interests

The authors have nothing to declare.

Acknowledgement

This work is supported by NIH grants R01 EB016089 R01 EB023963 R21A G060245. GO receives support from NIH grant K99 AG062230. MP is supported by NIH grants P41EB015909 and R01NS106292.

CRediT authorship contribution statement

Helge J. Zöllner: Software, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Visualization. **Georg Oeltzschner**: Conceptualization, Methodology, Software, Investigation, Formal Analysis, Writing – Review & Editing, Supervision, Funding acquisition. **Michal Považan**: Investigation, Data Curation, Writing- Review & Editing. **Steve C. N. Hui**: Software, Investigation, Writing – Review & Editing. **Sofie Tapper**: Investigation, Writing – Review & Editing. **Richard A. E. Edden**: Conceptualization, Formal Analysis, Writing – Review & Editing, Supervision, Project administration, Funding acquisition